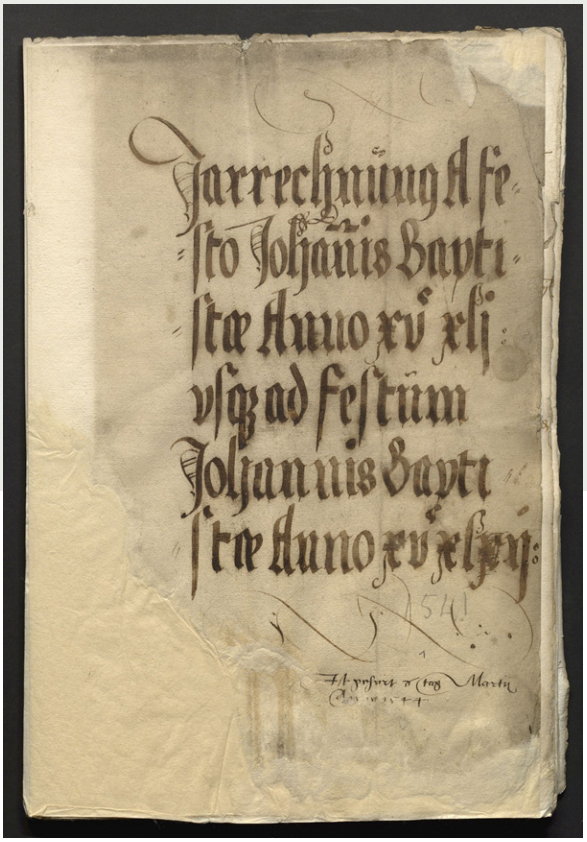
BASEL DIGITAL FRÜHE NEUZEIT WIRTSCHAFT

Reading an XML Text Like a Human with Semantic Web Technologies

Learning from the Basel City Accounts as Digital Edition

Georg Vogeler 26.05.2021



Historical texts carry information. Though this assertion will hardly surprise historians, they less often consider how much this information is influenced by the material, visual, and organisational context of the text. Which archive considered the text worth preserving? How is the text organised on the page? Who wrote the text? What material was used to create the text? Digital editing allows us to create representations of texts which take into account all these aspects.

The digital scholarly edition of the Basel city accounts 1535–1611, created in collaboration with Susanna Burghartz and teams from Basel and Graz, and published in 2015,[1] demonstrates both the feasibility and effectiveness of this digital approach to editing historical accounting records. It has become a major reference for digital scholarly editions of historical accounts.

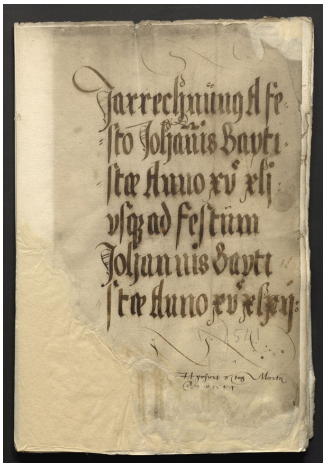


Fig. 1: Cover sheet of the annual account of the City of Basel 1541/42 (StABS, Finanz H, 98.1), Source.

Starting from the experience in the work with this historical object, this contribution reflects on technical solutions which can help to realise editions similar to the Basel city accounts. The edition uses a combination of two well-established technologies in the Digital Humanities: the Text Encoding Initiative (TEI[2]) XML markup for transcriptions, and the Resource Description Framework (RDF[3]), defined by the World Wide Web consortium (W3C), to publish the content of the accounts in the «Web of Data» or the «Semantic Web». RDF is based on globally unique identifiers in the form of web addresses (URI[4]) and a formal structure similar to simple textual statements: «subject, object, predicate». These «triples» can be read as propositions about the world. As such, RDF triples are useful for explicitly modelling the semantic content of historical editions.[5] I have even argued that this approach can be considered a specific type of scholarly edition, which I have called the «assertive edition».[6]

The edition of the Basel accounts was realised by transcribing the accounts and embedding mark-up with TEI/XML. The mark-up schema used is a customization of the standard TEI encoding with records-specific tags: individual entries (`r:e`) containing amounts of money (`r:b`), received or spent by the city, and summed in totals (`r:sum`); and with several statements aggregated with brackets (`r:klammer`). The entries are organised in rubrics referencing a taxonomy common to all seventy annual records.[7]

```
–<div ana="#bs_Mühlenungeld" xml:id="bs_Mühlenungeld-div-0">
   <head xml:id="d2e205" xml:space="preserve">Vom mülikorn ungellt</head>
   <metamark function="aggregate" rend="Klammer" spanTo="#bs_Mühlenungeld-total-1" target="#bs_Mühlenungeld-1"/>
  –<p ana="#bk_entry" rend="klammer" xml:id="bs_Mühlenungeld-1">
      Prima angaria
     –<seg ana="#bk_amount" rend="rb" xml:id="d2e212" xml:space="preserve">
       –<measure quantity="1110" type="currency" unit="lb">
          j
          <seg rend="super">m</seg>
          j
          <seg rend="super">c</seg>
          x lb
        </measure>
      </seg>
      <del xml:id="d2e221">viij ß</del>
   </p>
  –<p ana="#bk_entry" rend="klammer" xml:id="bs_Mühlenungeld-2">
      Secunda angaria
     –<seg ana="#bk_amount" rend="rb" xml:id="d2e226" xml:space="preserve">
       –<measure quantity="1102" type="currency" unit="lb">
          j
          <seg rend="super">m</seg>
          j
          <seg rend="super">c</seg>
          ij lb
        </measure>
```

Fig. 2: TEI encoding from https://gams.uni-graz.at/o:srbas.1541 preserves brackets, deletions, paleography of the number, etc.

These transcriptions are stored and disseminated in the *GAMS* application,[8] which splits them into a representation of the text in pure TEI and a representation of the accounting information in RDF during ingest into the system. The user can search and calculate with the RDF database and read the transcripts from the TEI.

```
–<rdf:Description rdf:about="http://gams.uni-graz.at/o:srbas.1541/sdef:TEI/get#bs_Mühlenungeld-1">
    <rdf:type rdf:resource="http://gams.uni-graz.at/rem/bookkeeping/#entry"/>
    <g2o:partOf rdf:resource="http://gams.uni-graz.at/o:srbas.1541"/>
    <bk:account rdf:resource="http://gams.uni-graz.at/o:srbas.konten#bs_Einnahmen"/>
    <bk:account rdf:resource="http://gams.uni-graz.at/o:srbas.konten#bs_StadtEinnahmen"/>
    <bk:account rdf:resource="http://gams.uni-graz.at/o:srbas.konten#bs_Mühlenungeld"/>
    <bk:mainAccount rdf:resource="http://gams.uni-graz.at/o:srbas.konten#bs_Mühlenungeld"/>
  –<bk:amount>
    –<rdf:Description rdf:about="http://gams.uni-graz.at/o:srbas.1541/sdef:TEI/get#d2e212">
        <bk:as rdf:resource="http://gams.uni-graz.at/rem/bookkeping/#i"/>
        <bk:num rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">266400</bk:num>
        <bk:unit rdf:resource="http://gams.uni-graz.at/rem/currencies/#d"/>
      </rdf:Description>
    </bk:amount>
    <bk:inhalt>Prima angaria jm jc x lb viij ß</bk:inhalt>
    <oa:hasTarget rdf:resource="http://gams.uni-graz.at/o:srbas.1541/sdef:Canvas/getJSON?context=fol.2r"/>
  </rdf:Description>
```

Fig. 3: The RDF representation of https://gams.uni-graz.at/o:srbas.1541 demonstrates the data structure of the entry https://gams.uni-graz.at/o:srbas.1541#bs_Mühlenungeld-1 including conversion of the amount into pennies and insertion into logical structure.

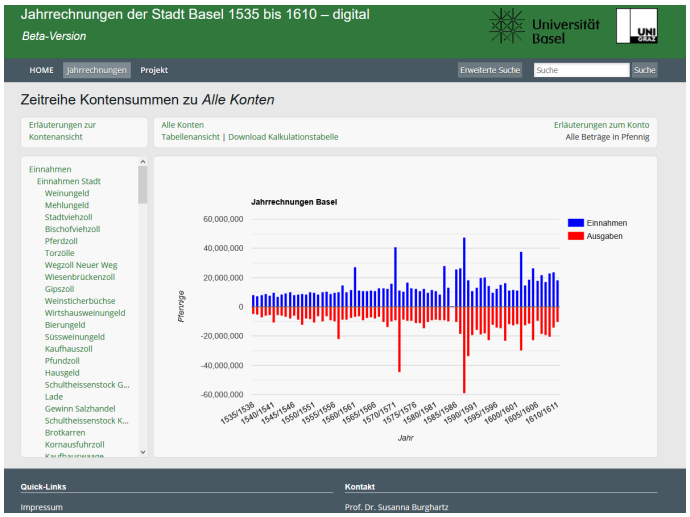


Fig. 4: Querying the RDF data of the Basel City Accounts can result in visualisations of income and expenses over time, Source.

This ingestion process is realised with the XML transformation language XSLT[9]. The script for converting the original encoding into RDF uses the structure of the original text, as represented in the XML structure of the digital document.[10] In the following I will explore how this specific process might be generalised. It addresses readers with an understanding of XML, TEI, RDF and OWL, although it will try to introduce most of the technical terminology at least by links to further information.

XML embeds hierarchically nested mark-up into the text, creating an ordered hierarchy of content objects.[11] Therefore, the data model of an XML text is both a sequence of words and annotations, and a tree of nested annotations. The generic semantic of sequence and nesting in XML annotations reflects human processing of text: humans try to sort sentences into blocks of information or into rhetorical superstructures like introductions, main arguments, and conclusions to contain arguments in paragraphs and chapters subsumed under headings and titles. Lists are a prototype for this organisation of text.[12] On the other hand, text has a sequence: a heading explains the following text; a conclusion is drawn from what is said before.

We can thus make use of this generic semantic of the XML meta-model in the extraction of structured information from XML encoded transcriptions. Both precedence in the annotation sequence and relative proximity to the root node (i.e. depth in the tree) can be interpreted as a sign that these nodes dominate the meaning of

nodes later in the text or deeper in the tree.[13] This can be used to extract RDF propositions from the XML mark-up. The necessary context to do this can be given by a schema of the vocabulary used for the RDF statements — the so-called 'TBox' of a formal ontology[14] (in contrast to the 'ABox' containing RDF statements on single entities) usually expressed in the Web Ontology Language 'OWL'[15]. This is based on two assumptions: first, that text gives the information necessary to understand other information before the occurrence of the latter; and, second, that hierarchy is used to reduce information redundancy, thus information given at a higher level is assumed to be applicable to all contained objects.

Currently, the data representation of the editions of the Basel accounts in the context of *GAMS* is realised via explicit rules of conversion. They are similar to those suggested by Torsten Schrade's *X-Triples*,[16] Hans-Jürgen Rennau's *RDFe*,[17] Herminio García's *ShExML*,[18] and most recently *eXGraphs*.[19] In *GAMS*, XSLT Stylesheets extract the RDF statements from the TEI. This is done by an extensive use of the TEI/XML `@ana` construct,[20] as it can be applied to any kind of TEI mark-up. The content of this attribute is connecting stand-off annotation to the embedded mark-up. It references URIs in a TBox of 'task ontologies' of the projects, i.e. formal ontologies for a single task. Some of these ontologies are in the process of becoming 'domain ontologies', i.e. covering the needs of several similar projects,[21] while some are in an earlier stage, e.g. legal cases.[22] The bookkeeping ontology developed in the context of the edition of the Basel accounts has grown into a proposal for a domain ontology in follow-up projects, for instance https://gams.uni-graz.at/o:depcha.bookkeeping. When these ontologies become more expressive, it is hoped that they will create constraints which can be systematically translated into XSLT/XPath expressions in a generalised way.

However, Web Ontology Language (OWL) is not sufficient to extract relationships between entities marked-up in the text that are not explicit: It can infer relationships from axioms. It cannot infer relationships from non-existent axioms, i.e. create constraints to be satisfied by the current resource alone. This is caused by the open world assumption under which OWL as defined by the W3C works. An OWL TBox states that a financial transaction consists of transfers; however, it does not state that a specific dataset has to fulfil this assertion. Under the open world assumption, the transfers of the transaction recorded in the current account could be stored elsewhere in the semantic web. Thus, we cannot infer an XSLT/XPath to be applied to the document from this TBox.

W3C's Shapes Constraint Language (SHACL)[23] is another approach to describe rule sets for RDF data. SHACL shapes have the capability to make the TBox more expressive, as they can test completeness of the ABox. A SHACL TBox of accounting records in Listing 1 includes formal definitions of the historical assumption that each transaction has at least two partners who exchange one commodity for an amount of money.

```
prefix sh: <http://www.w3.org/ns/shacl#>
prefix rdf:          <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix bk: <http://www.gams.uni-gaz.at/rem/ontology#>

bk:TransactionShape
          a sh:NodeShape ;
          sh:targetClass bk:Transaction ;
          sh:property [
                    sh:path bk:consistOf ;
                    sh:minCount 2 ;
                    sh:class bk:Transfer ;
                    sh:nodeKind sh:IRI ;
          ] ;
          sh:closed true ;
          sh:ignoredProperties ( rdf:type ) .
bk:TransferShape
          a sh:NodeShape ;
          sh:targetClass bk:Transfer ;
          sh:property [
                    sh:path bk:transfers ;
                    sh:Class bk:Measurable ;
                    sh:minCount 1 ;
          ] ;
          sh:property [
                    sh:path bk:from ;
                    sh:Class bk:Between ;
                    sh:minCount 1 ;
          ] ;
          sh:property [
                    sh:path bk:to ;
                    sh:Class bk:Between ;
                    sh:minCount 1 ;
          ]
          sh:closed true ;
          sh:ignoredProperties ( rdf:type ) .
```

Listing 1: SHACL shape examples for bookkeeping ontology.

The SHACL shapes can control the result of a XSLT conversion of the analytical semantics encoded in `@ana` attributes in an XML/TEI-mark up into RDF. But the relationships required in these shape definitions are hidden in the XML hierarchy:

```
<text>
  <head ana=”bk:between”>Horse retailer Inc.</head>
  <div>
    <head ana=”bk:from”>Hans</head>
    <p ana=”bk:transfer”><measure ana=”bk:transfers”>1 horse</measure>
<measure ana=”bk:amount”>100 $</measure></p>
  </div>
</text>
```

The XML contains local information about the transaction in the p-Element. The partners in the transaction are encoded in XML fragments before the transaction in question. To include them in the extracted RDF, the list of conditions in the TBox can be tested based only on the hierarchical and sequential assumptions of the XML

meta-model. The assumption that text gives information necessary to understand other information usually before the occurrence of the fragmentary information set and that hierarchy can help translating the SHACL definitions into XPath-expressions. The following expresses the requirement of having at least `one bk:from` object in a shape targeted at the class `bk:Transfer` with the class `bk:Between` as range:

```
bk:TransferShape
          a sh:NodeShape ;
          sh:targetClass bk:Transfer ;
          sh:property [
                    sh:path bk:from ;
                    sh:Class bk:Between ;
                    sh:minCount 1 ;
          ] .
```

The shape definition is converted into an XSLT template creating statements for each XML element, which map to the target classes via the `@ana` attribute:

```
<xsl:template match="*[@ana='bk:Transfer']">
<bk:Transfer ID="{@xml:id}"><!-- here the properties ...--></bk:Transfer>
</xsl:template>
```

This template creates properties for the entity according to the definition of the shape in SHACL. This results in the following XPath (in bold) to address the `bk:from` property when processing an XML element associated with the `bk:Transfer` class (again via the `@ana` attribute):

```
<xsl:template match="*[@ana='bk:Transfer']">
<bk:Transfer ID="{@xml:id}">
<xsl:for-each select="(descendant::*[@ana="bk:from" | ancestor-or-self::*[@ana="bk:from"][1] | preceding::*[@ana="bk:from"][1])">
  <bk:from><!-- here the content of the bk:from ... --></bk:from>
</bk:Transfer>
</xsl:template>
```

The result would still not fulfil the requirements of the SHACL shape for the following part of the shape definition:

```
bk:TransferShape
          a sh:NodeShape ;
          sh:targetClass bk:Transfer ;
          sh:property [
                    sh:path bk:to ;
                    sh:Class bk:Between ;
                    sh:minCount 1 ;
          ]
```

In this case, the conversion has to infer the `bk:to` from the given data. The SHACL shape tells the system that `bk:to` is necessary and it is an instance of the class `bk:Between`. With this information, the following XPath can be used to identify the `bk:to` property in the triple, when no explicit `bk:to` encoding is given:

```
(descendant::*[@ana="bk:Between" | ancestor-or-self::*[@ana="bk:Between"][1] | preceding::*[@ana="bk:Between"][1])
```

This method replaces the explicit conversion patterns of *xTriples, eXGraph* or *GAMS*-ToRDF with a formal description of the expected RDF output in SHACL, and infers from this a set of XSLT templates and XPaths. Applying them to the `@ana` attributes of the given TEI extracts the semantic structure and represents it in RDF.

The engine for generating XSLT from SHACL shapes has not yet been built. It would use SPARQL queries of the SHACL shapes to create XSLT templates that will match the target class of `@ana` attributes of the XML/TEI edition, thus creating the properties defined by the shape.

Compared to ad hoc conversion patterns, by using SHACL we gain an already-standardised method for defining patterns. A further advantage is that the resulting RDF can be validated against the SHACL shapes, thus keeping conversion and resulting RDF consistent.

There are no mechanisms yet to automatically create these formalisations, although SHACL provides the necessary formal definitions. However, assertive editions which are in the core interest of historians like Susanna Burghartz need these mechanisms. Digital Humanities has started to work on technologies helping to express description logic and constraints as it is formalised in OWL and SHACL. The single edition of the Basel accounts has triggered wide ranging conceptual and methodological considerations – and still serves as a best practice of scholarly editing of historical accounts with digital means.

Empfohlene Zitierweise/Suggested citation

Georg Vogeler: Reading an XML Text Like a Human with Semantic Web Technologies – Learning from the Basel City Accounts as Digital Edition. In: Tina Asmussen, Eva Brugger, Maike Christadler, Anja Rathmann-Lutz, Anna Reimann, Carla Roth, Sarah-Maria Schober, Ina Serif (Hg.): Materialized Histories. Eine Festschrift 2.0, 26/05/2021, https://mhistories.hypotheses.org/?p=783.

Abstract

The digital scholarly edition of the Basel City Accounts has been successful in its combination of TEI/XML encoded transcriptions with RDF representation of the historical content of the accounts. The contribution discusses the use of the RDF rule language SHACL to define a formalism to convert the implicit structure of the XML encoding into explicit RDF statements. It suggests to use precedence and dominance of XML elements as indicators for the implicit information. SHACL can serve as the formalisation of constraints. used to extract XPath expressions, which include missing explicit information from elements in the ancestor or preceding axis. With this method, a formal description of the target data model can be used to generate XSLTs converting the TEI/XML into RDF/XML.

1 Susanna Burghartz (ed.): Jahrrechnungen der Stadt Basel 1535–1610. Basel/Graz 2015, http://gams.uni-graz.at/context:srbas [14.01.2021].
2 https://tei-c.org/Guidelines/P5/ [14.01.2021].
3 https://www.w3.org/RDF/ [14.01.2021].
4 https://en.wikipedia.org/wiki/Uniform_Resource_Identifier [14.01.2021].
5 Georg Vogeler: Digital Edition of Archival Material – Machine Access to the Content. On the Role of Semantic Web Technologies in Digital Scholarly Editions, in: Christel Loubet (ed.): Digitizing Medieval Sources – L'édition en ligne de documents d'archives médiévaux. Challenges and Methodologies – Enjeux, méthodologie et défis. Turnhout 2020, pp. 37–56, DOI: 10.1484/M.ARTEM-EB.5.117327 [14.01.2021].
6 Georg Vogeler: The 'Assertive Edition'. On the consequences of digital methods in scholarly editing for historians, in: International Journal of Digital Humanities, 1/2 (2019), pp. 309–322, https://doi.org/10.1007/s42803-019-00025-5 [14.01.2021].
7 http://gams.uni-graz.at/context:srbas?mode=projekt#die-digitale-edition [14.01.2021].
8 Johannes Stigler, Elisabeth Steiner: GAMS – An Infrastructure for the Long-Term Preservation and Publication of Research Data from the Humanities, in: Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71/1 (2018), pp. 207–216. DOI: 10.31263/voebm.v71i1.1992 [14.01.2021].
9 https://en.wikipedia.org/wiki/XSL_Transformation [14.01.2021].
10 http://gams.uni-graz.at/archive/objects/cirilo:srbas/datastreams/TORDF/content [14.01.2021].
11 Steven J. DeRose, David G. Durand et al.: What is Text, really? in: Journal of Computing in Higher Education, 1/2 (1990), pp. 3–26.
12 Jack Goody: What's in a List? in: id.: The Domestication of the Savage Mind. Cambridge 1977, pp. 74–111.
13 Jennifer Tennison: Overlap, Containment and Dominance, 06.12.2008, https://www.jenitennison.com/2008/12/06/overlap-containment-and-dominance.html [14.01.2021].
14 https://en.wikipedia.org/wiki/Ontology_(information_science) [14.01.2021].
15 https://www.w3.org/OWL/ [14.01.2021].
16 Akademie der Wissenschaften und der Literatur Mainz, Digital Academy: XTriples, 2012 https://xtriples.lod.academy/index.html [14.01.2021]; Akademie der Wissenschaften und der Literatur Mainz. EXGraphs. LOD.Academy (2019), https://lod.academy/site/tools/digicademy/exgraphs [14.01.2021]; Torsten Schrade: digicademy/xtriples: 1.4.0, 25.03.2019, DOI: 10.5281/zenodo.2604986 [14.01.2021].
17 Hans-Jürgen Rennau: RDFe – Expression-Based Mapping of XML Documents to RDF Triples, in: XML Prague 2019. Conference Proceedings. Prague 2019, pp. 381–404. https://archive.xmlprague.cz/2019/files/xmlprague-2019-proceedings.pdf#page=393 [14.01.2021].
18 http://shexml.herminiogarcia.com/ [14.01.2021]; Herminio Garcia-Gonzalez, Daniel Fernandez-Alvarez et al.: ShExML: An Heterogeneous Data Mapping Language Based on ShEx, in: Philipp Cimiano, Olivier Corby (ed.): EKAW-PD 2018 Posters and Demonstrations at EKAW 2018. Proceedings of the EKAW 2018. Posters and Demonstrations Session co-located with 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018). Nancy, France, November 12–16,(2018, CEUR-Workshops 2262), pp. 9–12, http://ceur-ws.org/Vol-2262/ekaw-poster-08.pdf [14.01.2021].
19 Akademie der Wissenschaften und der Literatur Mainz, & Digital Academy: XTriples (2012), https://xtriples.lod.academy/index.html [14.01.2021]; Akademie der Wissenschaften und der Litera-

tur Mainz: EXGraphs. LOD.Academy (2019), https://lod.academy/site/tools/digicademy/exgraphs [14.01.2021].

20 https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.global.analytic.html [14.01.2021].

21 Christopher Pollin, Georg Vogeler: DEPCHA – Digital Edition Publishing Cooperative for Historical Accounts. Graz 2018, http://gams.uni-graz.at/context:depcha [14.01.2021].

22 Urfehdebücher der Stadt Basel – digitale Edition, ed. by Susanna Burghartz, Sonia Calvi et al., Basel/Graz 31.01.2017, http://gams.uni-graz.at/ufbas [14.01.2021]; Christopher Pollin, Georg Vogeler: Semantically Enriched Historical Data: Drawing on the Example of the Digital Edition of the «Urfehdebücher der Stadt Basel», in: Alessandro Adamou, Enrico Daga et al. (ed.): Workshop on Humanities in the Semantic Web. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II). Co-Located with 16th International Semantic Web Conference (ISWC 2017), (CEUR Workshop Proceedings 2014). Budapest 2017, pp. 27–32, http://ceur-ws.org/Vol-2014/paper-03.pdf [14.01.2021].

23 Holger Knublauch, Dimitris Kontokostas: Shapes Constraint Language (SHACL), W3C Recommendation 20.07.2017, https://www.w3.org/TR/shacl/ [14.01.2021].