Journal of Catalan Intellectual History (JOCIH)
Online ISSN 2014-1564 DOI: 10.1515/jocih-2016-0014
https://www.degruyter.com/view/j/jocih

Received 22/02/2017 / Accepted 02/04/2017

The Aberdeen Burgh Records of 1398–1531 and the Semantic Web

Anna D. Havinga[*] (University of Bristol, Department of German)

Anna.Havinga@bristol.ac.uk https://orcid.org/0000-0002-6470-423X

Adam Z. Wyner[**] (University of Aberdeen, Department of Computing Science)

azwyner@abdn.ac.uk https://orcid.org/0000-0002-2958-3428

Abstract. This paper provides an overview of two text analytic projects on the Aberdeen burgh records, which are legal records of the city of Aberdeen, Scotland. These records contain detailed information about a range of activities in the city and their legal treatment. The projects cover the periods 1398–1511 (Law in the Aberdeen Council Registers project – LACR) and 1530–1531 (A Text Analytic Approach to Rural and Urban Legal Histories project – TAHL). The completed TAHL project annotated a selected corpus with rich semantic information for the purpose of facilitating historical research by querying and extracting data from across the corpus. The LACR project, which is ongoing, focuses on transcribing the first eight volumes of the Aberdeen burgh records (1398–1511) into the Text Encoding Initiative's standard, thus making the text machine-readable. This project lays the foundation for further analysis and enrichment of the corpus.

Keywords: Digital Humanities, Semantic Web, Legal History

1 Introduction

The council registers of Aberdeen, Scotland are the earliest and most complete body of town (or burgh) council records in Scotland, running nearly continuously from 1398 to the present. Few cities in the United Kingdom or in Western Europe rival Aberdeen's burgh registers in historical depth and completeness. In July 2013, UNESCO UK recognised the register's volumes from 1398 to 1509 as being of outstanding historical importance to the UK (Aberdeen City Council, 2013). The registers offer a detailed legal view into one of Scotland's principal burghs, casting light on administrative, legal, and commercial activities as well as daily life. The registers include the elections of office bearers, property transfers, regulations of trade and prices, references to crimes and subsequent punishment, matters of public health, credit and debt, cargoes of foreign vessels, tax and rental of burgh lands,

[*] Dr Anna D. Havinga is Lecturer in Sociolinguistics at the University of Bristol. She is particularly interested in language standardisation, language contact, and language change. Her monograph on *Invisibilising Austrian German* will appear in 2018 (see https://www.degruyter.com/view/product/489082).

[**] Dr Adam Z. Wyner is Lecturer in Computing Science at the University of Aberdeen. His research interests range over topics in Artificial Intelligence and Law, including argumentation, natural language processing, legal reasoning, historical text processing, and ontologies. A recent paper is: O. Shulayeva, A. Siddharthan, and A. Wyner. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126, 2017.

woods, and fishing. The entries thus present the burgh's relationships with the countryside and countries around the North Sea.

To make this historical resource available to a wider audience, the National Records of Scotland and Aberdeen City and Aberdeenshire Archives collaborated to image the volumes digitally up to 1511 and made them available on the internet.[1] However, the images of scribal records are inaccessible to all but a few scholars since they are handwritten in Latin and Middle Scots (mainly).

To address this issue, an initial pilot project at the University of Aberdeen's Research Institute of Irish and Scottish Studies (RIISS) has transcribed 100 pages of the records from the period 1530–1531, translated the Latin and Middle Scots, and provided a web-accessible database application; the application allows users to query the database for locations and names of individuals, and returns the textual portions that contain those names and locations.[2]

However, the initial pilot project does not make use of any current techniques or technologies from digital humanities, text analysis, or the Semantic Web to facilitate understanding of and access to the records. To begin to apply such techniques and technologies, a subsequent, completed pilot project, *Text Analytic Approach to Urban and Rural Histories*, applied tools to enrich the transcribed 100 pages with semantic information, which could be queried, reprocessed, and linked. Following this, a further larger project, *Law in the Aberdeen Council Registers*, which is in progress (March 2016 to February 2019), endeavours to transcribe the first eight volumes from 1398–1511 in a Text Encoding Initiative (TEI)-compliant form, making the texts machine-readable.

The following sections outline these two latter projects in terms of goals, techniques, and outcomes. The paper concludes with some observations of current progress and future opportunities.

2 A Text Analytic Approach to Rural and Urban Legal Histories (TALH)

In this section, we will give an overview of TALH, which applied text analytic tools to a sample of the Aberdeen Council Registers.[3] The underlying motivation of TALH was to facilitate research with the help of text analytic tools. In other words, how can we translate the questions that scholars have into text analytic queries, which will allow the search and retrieval of information from across a large corpus. For example, for legal historians, the burgh registers are an opportunity to study source materials concerned with the law and community concerning questions such as:

- What legal roles in jurisdictions do individuals perform?
- What are the social and legal networks?
- How do social and legal concepts evolve?
- What does the historical record say about resource management and conflict?

While traditional historical methods applied to archival material has served well enough, it is costly, slow, and does not allow analysis of the information in large and complex corpora. Moreover, some of the questions above are relational, e.g. relations of individuals in legal roles, which are difficult to track across a large corpus. With text analytic support, legal historians can query a corpus and receive data either in context or extracted.

To facilitate querying the corpus, the text was enriched with semantic annotations. The General Architecture for Text Engineering (GATE) framework was used to add the annotations to the text. GATE is used for language

[1] http://www.scotlandsplaces.gov.uk/digital-volumes/burgh-records/aberdeen-burgh-registers/
[2] http://www.abdn.ac.uk/riiss/about/database-159.php
[3] http://www.dotrural.ac.uk/talh/

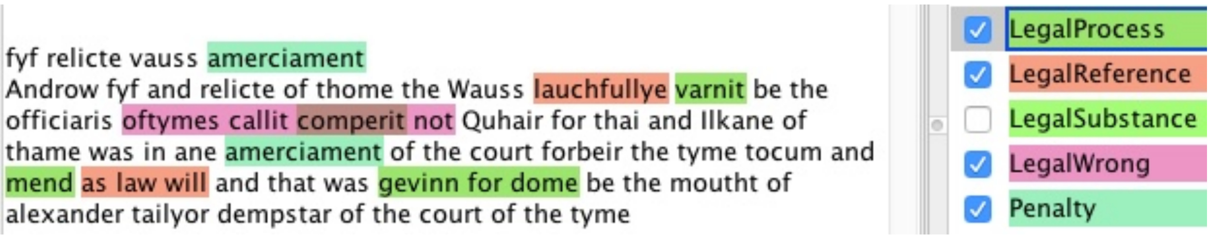
engineering applications, supports efficient, robust text processing, and has been applied in many large text processing projects (Cunningham et al., 2002). It is a highly scalable, open source, desktop application written in Java. GATE provides a user interface that allows professional linguists and text engineers to bring together a wide variety of natural language processing tools and apply them to a set of documents. The tools are formed into a pipeline of natural language processors. Once a GATE pipeline has been applied, we can view the annotations either in situ or query them using GATE's ANNIC (ANNotations In Context) corpus indexing and querying tool.

For our purposes, the key processing stage was *lookup*, wherein textual passages in the corpus were automatically matched with terms on a list, then assigned an annotation; e.g. a token term *burgi* is annotated *LegalBody*. Similarly, tokens such as *common council*, *curia*, *guild court*, and others were all annotated *LegalBody*. The list items and their annotations were manually constructed. Thus, the *lookup* stage is used to automatically annotate related terms (e.g. *burgi* and *guild court*) with the same annotation (e.g. *LegalBody*); in this way, annotations serve as conceptual covers for tokens. We had lists that provide a range of semantic concepts for entities and relations, such as the ones listed in Table 1.

Table 1: *Sample of semantic concepts with examples*

Annotations	Examples
LegalBody	burgi, common council
LegalConcept	gude faith
LegalRole	Archbishop, Bailie
Offence	barganyng, tulyheing
Office	alderman, burgess
RegisterEntry	Bailie Court, Ordinance

The annotations were represented as XML tags (so machine readable) or (using XSLT) visually by coloured highlighting, making the annotations immediately apparent. Fig. 1 illustrates the menu of annotations on the right and the marked-up passages on the left.

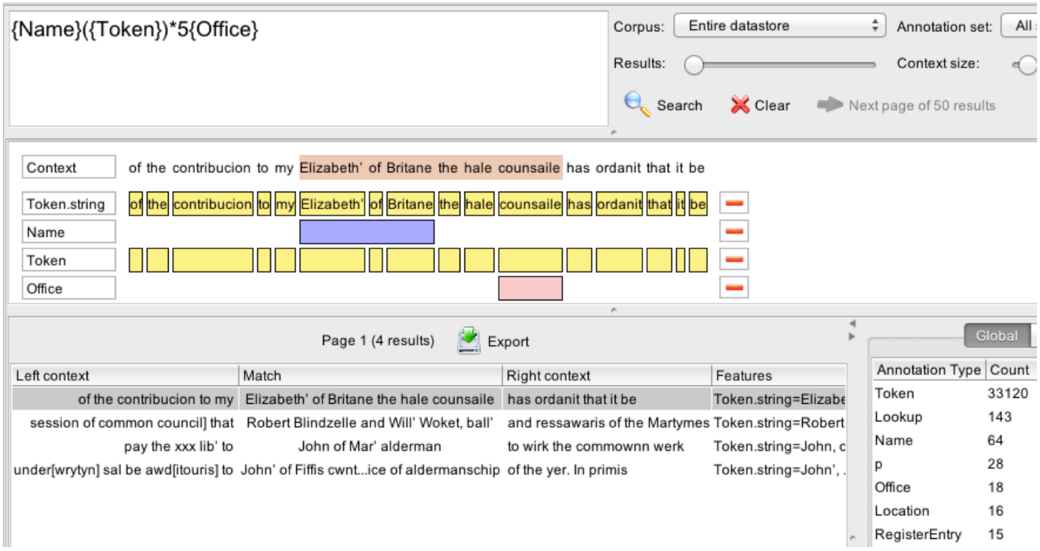


fyf relicte vauss amerciament
Androw fyf and relicte of thome the Wauss lauchfullye varnit be the
officiaris oftymes callit comperit not Quhair for thai and Ilkane of
thame was in ane amerciament of the court forbeir the tyme tocum and
mend as law will and that was gevinn for dome be the mouth of
alexander tailyor dempstar of the court of the tyme

☑ LegalProcess
☑ LegalReference
□ LegalSubstance
☑ LegalWrong
☑ Penalty

Fig. 1: *Example of annotations and marked-up passages in GATE*

We used the ANNIC tool to index and query a database of annotated text. Searching in the corpus for single annotations returns all strings that are annotated with the search annotation along with their context and source document. Complex queries can also be formed. A query and a sample result appear in Fig. 2, where the query finds all sequences of annotated text where the first string is annotated with *Name*, followed by zero to five other *Tokens*, followed by a string with an *Office* annotation:



```
{Name}({Token})*5{Office}
```

Corpus: Entire datastore
Annotation set: All
Results:
Context size:

Context: of the contribucion to my Elizabeth' of Britane the hale counsaile has ordanit that it be

Token.string: of | the | contribucion | to | my | Elizabeth' | of | Britane | the | hale | counsaile | has | ordanit | that | it | be

Left context	Match	Right context	Features
of the contribucion to my	Elizabeth' of Britane the hale counsaile	has ordanit that it be	Token.string=Elizabe
session of common council] that	Robert Blindzelle and Will' Woket, ball'	and ressawaris of the Martymes	Token.string=Robert
pay the xxx lib' to	John of Mar' alderman	to wirk the commownn werk	Token.string=John, c
under[wrytyn] sal be awd[itouris] to	John' of Fiffis cwnt...ice of aldermanschip	of the yer. In primis	Token.string=John', .

Annotation Type	Count
Token	33120
Lookup	143
Name	64
p	28
Office	18
Location	16
RegisterEntry	15

Fig. 2: *Example of query and sample result in GATE*

The search returned four candidate structures. The extract identifies a relation between an individual and their office. Similar relational queries can be made about other aspects of the text. With the query language, we can search for any number of annotations in the corpus in any order; the tool allows incremental refinement of searches, resulting in a highly interactive way to examine the semantic content of the texts. Thus, a range of semantic patterns can be identified that would otherwise be very hard to detect or extract.

Such an approach can ground multidisciplinary investigations of historical societies in large-scale textual sources of information, providing interpretable material on topics such as elites and social practice, relations between social classes and land, urban and rural development, and natural resource management. The text analytic approach also makes applicable a range of social web-mining approaches on historical text.

The project delivered offset and inline XML files, the latter queryable with XPATH. In addition, the XML was converted to RDF triples, allowing SPARQL queries as well as linking TALH materials to external resources. However, given the limited resources, there were only limited experiments with querying and linking the representations (Wyner et al., 2014).

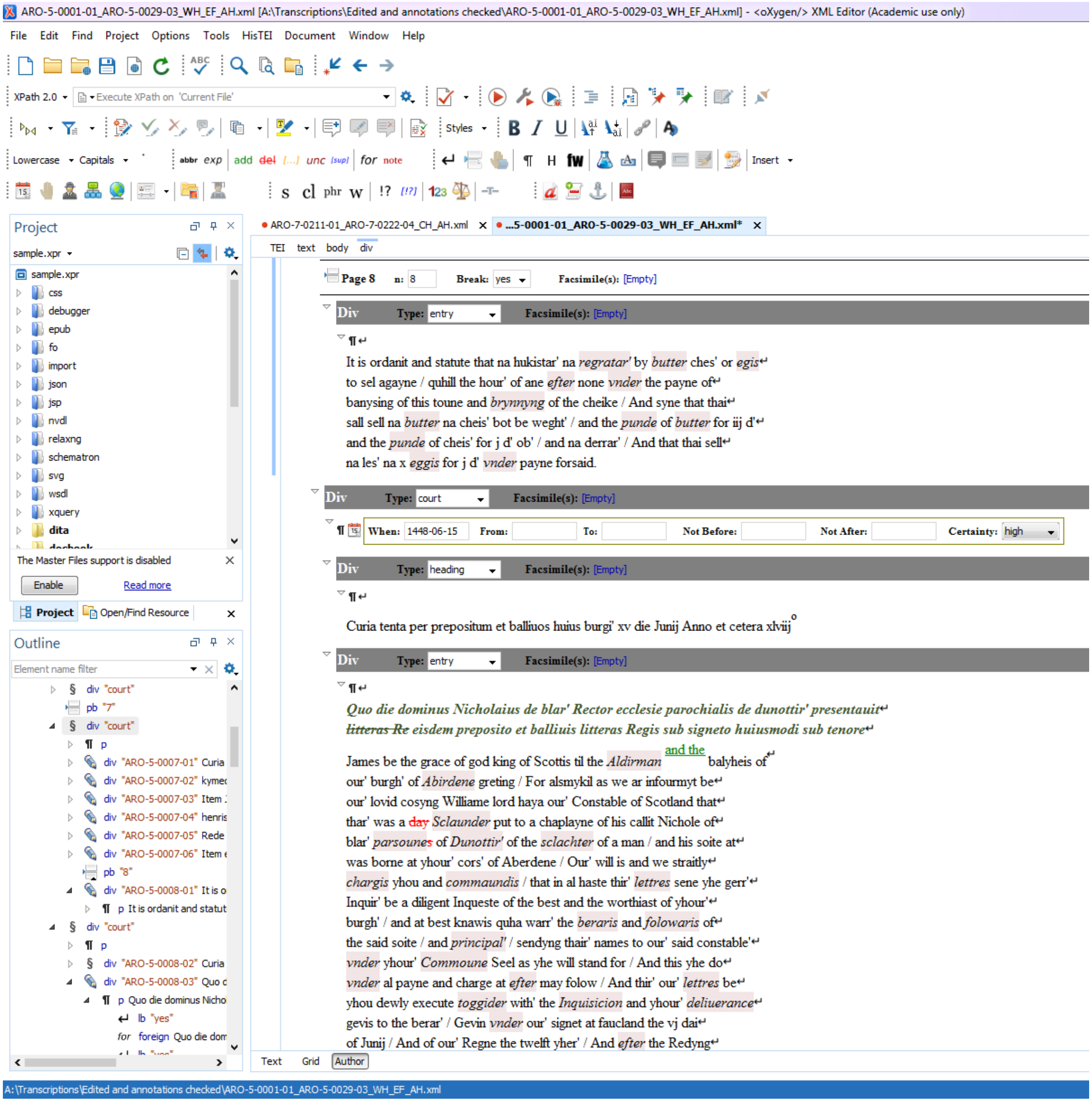
3 Law in the Aberdeen Council Registers (LACR)

Although the TALH project was mainly aimed at facilitating academic research, the main aim of the LACR project[4] is to make the Aberdeen Council Registers more widely accessible. To do so, the goal of the LACR project is to transcribe the first eight volumes of the handwritten records, covering the period 1398 to 1511. In the following, we describe some of the tools, workflow, and output of the LACR project.

The semi-diplomatic transcriptions are compliant with Text Encoding Initiative (TEI) guidelines (version P5) (TEI Consortium, 2016). The Oxygen XML editor is used with the add-on HisTEI to produce these transcriptions. HisTEI renders the toolbars in the author view of Oxygen to supply encoding buttons that are particularly useful for the transcription of historical documents, such as buttons for <del> and <expan> elements (Olson, 2014). These buttons and their shortcut keys allow for quick and easy mark-ups in a word processor-like view, which is particularly helpful for people with less experience in encoding languages. By the click of a button, the

[4] https://aberdeenregisters.org/

corresponding XML tag is inserted in the Oxygen text view and visualised in a particular font and/or colour in the author view (see Fig. 3Fig. 3).


ARO-5-0001-01_ARO-5-0029-03_WH_EF_AH.xml [A:\Transcriptions\Edited and annotations checked\ARO-5-0001-01_ARO-5-0029-03_WH_EF_AH.xml] - <oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools HisTEI Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Styles
Lowercase Capitals
abbr exp add del [...] unc [sup] for note
Insert
s cl phr w !? [!?] 123
Project
sample.xpr
sample.xpr
css
debugger
epub
fo
import
json
jsp
nvdl
relaxng
schematron
svg
wsdl
xquery
dita
The Master Files support is disabled
Enable
Read more
Project
Open/Find Resource
Outline
Element name filter
§ div "court"
pb "7"
§ div "court"
¶ p
div "ARO-5-0007-01" Curia
div "ARO-5-0007-02" kymec
div "ARO-5-0007-03" Item
div "ARO-5-0007-04" henris
div "ARO-5-0007-05" Rede
div "ARO-5-0007-06" Item
pb "8"
div "ARO-5-0008-01" It is o
¶ p It is ordanit and statut
§ div "court"
¶ p
div "ARO-5-0008-02" Curia
div "ARO-5-0008-03" Quo d
¶ p Quo die dominus Nicho
lb "yes"
for foreign Quo die dom
ARO-7-0211-01_ARO-7-0222-04_CH_AH.xml
...5-0001-01_ARO-5-0029-03_WH_EF_AH.xml*
TEI text body div
Text Grid Author
A:\Transcriptions\Edited and annotations checked\ARO-5-0001-01_ARO-5-0029-03_WH_EF_AH.xml

Page 8 n: 8 Break: yes Facsimile(s): [Empty]

Div Type: entry Facsimile(s): [Empty]

¶

It is ordanit and statute that na hukistar' na *regratar'* by *butter* ches' or *egis*
to sel agayne / quhill the hour' of ane *efter* none *vnder* the payne of
banysing of this toune and *brynnyng* of the cheike / And syne that thai
sall sell na *butter* na cheis' bot be weght' / and the *punde* of *butter* for iij d'
and the *punde* of cheis' for j d' ob' / and na derrar' / And that thai sell
na les' na x *eggis* for j d' *vnder* payne forsaid.

Div Type: court Facsimile(s): [Empty]

¶ When: 1448-06-15 From: To: Not Before: Not After: Certainty: high

Div Type: heading Facsimile(s): [Empty]

¶

Curia tenta per prepositum et balliuos huius burgi' xv die Junij Anno et cetera xlviij°

Div Type: entry Facsimile(s): [Empty]

¶

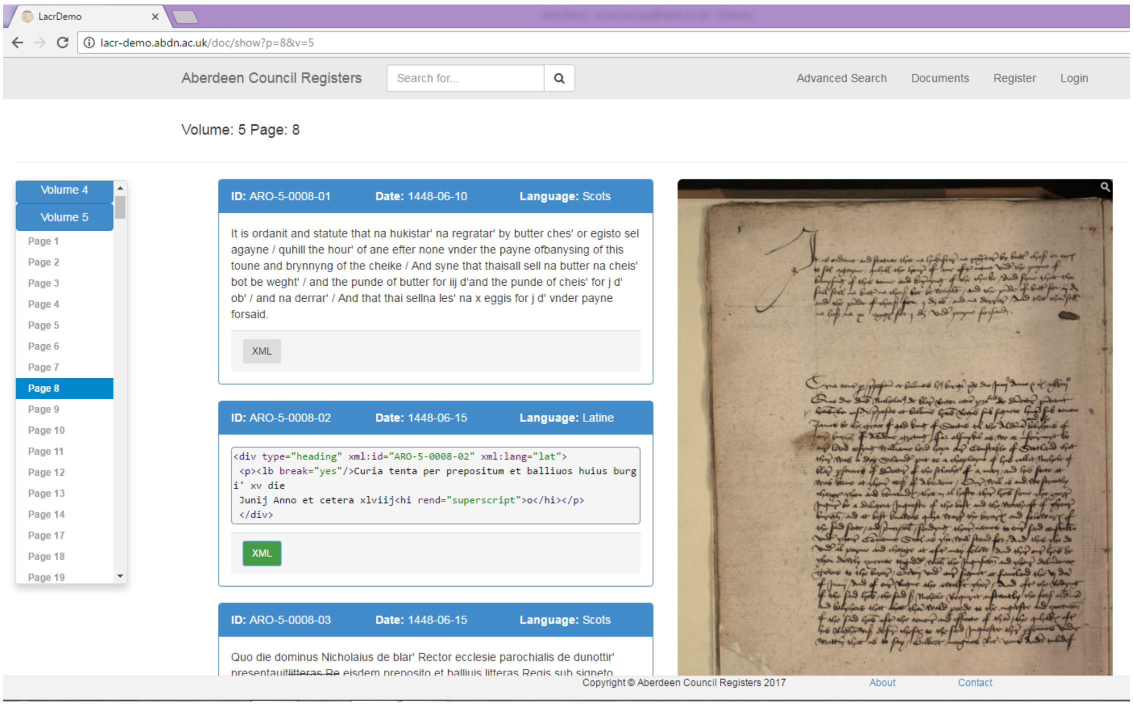
Quo die dominus Nicholaius de blar' Rector ecclesie parochialis de dunottir' presentauit
~~*litteras Re*~~ *eisdem preposito et balliuis litteras Regis sub signeto huiusmodi sub tenore*

James be the grace of god king of Scottis til the *Aldirman* and the balyheis of
our' burgh' of *Abirdene* greting / For alsmykil as we ar infourmyt be
our' lovid cosyng Williame lord haya our' Constable of Scotland that
thar' was a ~~day~~ *Sclaunder* put to a chaplayne of his callit Nichole of
blar' *parsoune*~~*s*~~ of *Dunottir'* of the *sclachter* of a man / and his soite at
was borne at yhour' cors' of Aberdene / Our' will is and we straitly
chargis yhou and *commaundis* / that in al haste thir' *lettres* sene yhe gerr'
Inquir' be a diligent Inqueste of the best and the worthiast of yhour'
burgh' / and at best knawis quha warr' the *beraris* and *folowaris* of
the said soite / and *principal'* / sendyng thair' names to our' said constable'
vnder yhour' *Commoune* Seel as yhe will stand for / And this yhe do
vnder al payne and charge at *efter* may folow / And thir' our' *lettres* be
yhou dewly execute *toggider* with' the *Inquisicion* and yhour' *deliuerance*
gevis to the berar' / Gevin *vnder* our' signet at faucland the vj dai
of Junij / And of our' Regne the twelft yher' / And *efter* the Redyng

Fig. 3: *Transcription with XML mark-up in Oxygen author view, using the HisTEI add-on*

After the initial transcription and annotation of the original texts by two research assistants, their transcriptions are checked and edited by the project manager. The XML annotations are then checked again separately by a research fellow on the team. A final inspection of the transcriptions and annotations is carried out once the initial transcription process is finished. These multiple checks, while being rather time-consuming, guarantee quality control.

The main outcome of the LACR project will be a versatile, accurate, TEI-compliant transcription of the first eight volumes of the Aberdeen Council Registers, available on an online platform. A prototype of this platform has been created by Computing Science students from the University of Aberdeen. Apart from hosting the transcriptions and images of the original text, which users can browse through, this platform allows users to search the text with respect to variable strings as well as XML annotations. The transcriptions of the original text are displayed as plain text by default, but the XML annotations can be viewed by clicking the XML button (see Fig. 4).



Volume: 5 Page: 8

ID: ARO-5-0008-01 **Date:** 1448-06-10 **Language:** Scots

It is ordanit and statute that na hukistar' na regratar' by butter ches' or egisto sel agayne / quhill the hour' of ane efter none vnder the payne ofbanysing of this toune and brynnyng of the cheike / And syne that thaisall sell na butter na cheis' bot be weght' / and the punde of butter for iij d'and the punde of cheis' for j d' ob' / and na derrar' / And that thai sellna les' na x eggis for j d' vnder payne forsaid.

ID: ARO-5-0008-02 **Date:** 1448-06-15 **Language:** Latine

```
<div type="heading" xml:id="ARO-5-0008-02" xml:lang="lat">
 <p><lb break="yes"/>Curia tenta per prepositum et balliuos huius burg
i' xv die
 Junij Anno et cetera xlviij<hi rend="superscript">o</hi></p>
 </div>
```

ID: ARO-5-0008-03 **Date:** 1448-06-15 **Language:** Scots

Quo die dominus Nicholaius de blar' Rector ecclesie parochialis de dunottir' presentauit~~litteras Re~~ eisdem preposito et balliuis litteras Regis sub signeto

Fig. 4: *LACR Prototype Platform, document view*

The search functionality of the platform encompasses both simple and advanced searches based on XQuery.[5] This allows users to carry out complex queries, which can be filtered by date, volume, language, and any other available XML annotation. The issue of spelling variation, which is very common in this period, is resolved by an adjustable filter for spelling variants using Elasticsearch.[6] If this filter is, for example, set to one spelling variant and the corpus is searched for *David*, the results will include *David* as well as *Dauid*. The search functionality also allows the use of regular expressions. The query \b[burg]\w*\b (word boundary – *burg* – followed by 0 to an unlimited number of word characters – word boundary), for example, will not only list all instances of *burg* but also *burgi*, *burges*, etc.

In the LACR project, the deliverables are essentially the files of the TEI-compliant XML transcriptions, which will be available via a website platform hosted by the Aberdeen City Archives, our project partners. The platform will allow users to search the records, view results, and download the XML files, enhancing the accessibility and academic use of the transcriptions. The annotations will be adjustable to suit the needs of researchers. The medieval manuscripts will thus not only be easily accessible for the general public but also searchable and adaptable for research purposes.

4 Discussion and Future Work

Both the TALH and the LACR project contribute to the application of language technologies for cultural heritage and the humanities. The TEI-compliant transcription of the Aberdeen Council Register is just the first step in enhancing online accessibility to these records, both for the general public and for scholars. As shown through the TALH project, semantic annotations allow for multidisciplinary investigations of corpora. Applying similar semantic annotations to the first eight volumes of the Aberdeen Council Registers would open a wealth of possibilities for historic research. Avenues for future projects also include linking the Aberdeen Council Registers to other local, regional, national, and international legal records. In addition, the Burgh records can make use of

[5] https://www.w3.org/XML/Query/
[6] https://www.elastic.co/

dictionaries, such as the Dictionary of the Scots Language,[7] either by linking to dictionary entries or by importing definitions into the text as annotations. The TALH project sampled some location names and mapped them to Aberdeen city maps; this could be extended over the LACR project, using Geographic Information System (GIS) data. Finally, the data from LACR could be party to the CLARIN network,[8] which would facilitate access to the LACR corpus from across the European network and application of CLARIN resources to the corpus.

Acknowledgments. The Law in the Aberdeen Council Registers project is funded by a Leverhulme Trust Research Project Grant. The project team comprises Dr Jackson Armstrong (Principal Investigator and Project Director), Dr Edda Frankot (Editorial Research Fellow and Project Manager), Dr Anna D. Havinga (former Text Enrichment Research Fellow), Dr Claire Hawes (Research Assistant), Dr William Hepburn (Research Assistant), Dr Andrew Mackillop (Co-Investigator), Dr Andrew Simpson (Co-Investigator), Dr Adam Wyner (Co-Investigator), and Phil Astley (Project Partner).

5 References

Aberdeen City Council: *Aberdeen's medieval burgh records recognized by UNESCO*, 2013 https://www.aberdeencity.gov.uk/CouncilNews/ci_cns/pr_archivesUNESCO_090713.asp [accessed 20/02/2017].

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL ·02)*, Philadelphia, USA: ACL Press, 2002, pp. 168-175.

Wyner, A., Armstrong, J., Mackillop, A., Astley, P.: Text Analysis of Aberdeen Burgh Records 1530-1531. In: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Gothenburg, Sweden: ACL Press, 2014, pp. 95-99.

TEI Consortium, eds.: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Version 3.1.0]. [Last modified 15/12/2016]. http://www.tei-c.org/Guidelines/P5/ [accessed 20/02/2017].

Olson, M.: *HisTEI. A framework for Oxygen XML Editor allowing researchers to transcribe historical documents in TEI*. 2014. http://www.histei.info/p/home.html [accessed 20/02/2017].

[7] http://www.dsl.ac.uk/

[8] https://www.clarin.eu/