USING COMPUTER ARTIFICIAL INTELLIGENCE TO TRANSCRIBE AND TRANSLATE ANCIENT ENGLISH LEGAL TEXTS

THE AIM OF THE PROJECT

The aim of the project is to use computer artificial intelligence techniques ("AI") to produce transcriptions of original ancient legal texts and to publish them online. Many such texts are not in English (usually mediaeval legal Latin) and, in those cases, part of the secondary process would be to produce translations into English from the transcriptions. Some transcriptions and translations have already been produced manually, notably by the Selden Society (www.selden-society.qmw.ac.uk), but there remain literally millions of pages of unprocessed texts. Transcribing and translating requires specialist expertise. Even if this could be applied to the task, it is likely to take many, many years. By using AI, it ought to be possible, after the initial setup, to reduce significantly the requirement for human specialist expertise and also to reduce the processing time by a major factor. By using deep-learning techniques, the computer should improve its skill and accuracy (and perhaps speed) as it is provided recursively with an ever increasing amount of training material (i.e. completed and checked texts).

THE JUSTIFICATION FOR THE PROJECT

Providing easy access to a huge body of previously inaccessible information will be immensely valuable to legal historians, social historians, archivists, genealogists and local historians, among others.

ESTABLISHING THE PROJECT

It is important that the project has well-respected credentials in order that both potential contributors and end-users have confidence in it. Since one of the major source locations, The National Archives, is located in London, it seems appropriate that the project should also be based in London, probably either at a centre of excellence for AI, such as Google's DeepMind (deepmind.com) or at an institution such as the University of London, Queen Mary College or Imperial College. These organisations should also have access to, or be able to work with, the specialist human experts and have access to the necessary computer resources to operate the project.

Consideration should be given to inviting involvement from suitable organisations, such as The National Archives, The British Library and universities, libraries and archives holding collections of suitable source material or having expertise that may be of assistance to the project.

The project is multi-disciplinary, and will require input not only from experts in computer-based deep learning but also from experts in legal history, palaeography and mediaeval legal Latin. Might it be a suitable PhD project ?

FUNDING THE PROJECT

It is assumed that the end product of the project will be freely accessible online or at least funded ultimately by access fees. Free access, such as is given by the Old Bailey Online project (www.oldbaileyonline.org) would perhaps be the ideal. It will therefore appear necessary to obtain funding through one or more grants or, perhaps through crowd-funding.

The project may need to be costed, unless it can be open-endedly sponsored.

HUMAN RESOURCES

Besides the human expertise identified above, the project will need to be managed. Additionally, there will be a need for moderators and/or editors to validate the computer output (see below).

THE SOURCES

The handwritten source texts for the English legal system exist from about the 13th century and continue until the common use of printing in the 18th century. Most are held in The National Archives, but there are collections in other archives, and in major libraries. Part of the project would be to locate as many such sources as possible. In preparation for computer processing, machine-readable images of the sources would have to be acquired by photography or scanning ("digitization").

The physical location of all sources would have to be recorded in such a way that a complete audit trail exists and can be easily accessed from the source to the end result.

IMAGING THE TEXTS

It is envisaged that the imaging process would *not* form part of the project, other than in exceptional cases. This is prompted by the Anglo-American Legal Tradition project ("AALT") (http://aalt.law.uh.edu/). According to their website, as at August 2015 they had digitized 9,250,000 frames of historical material from documents from mediaeval and early modern England from The National Archives. The images are freely accessible from the AALT website. I have been in contact with the head of AALT, Professor Robert C. Palmer (Cullen Professor of Law and History at the University of Houston, USA). He has informally indicated to me that AALT is unlikely to undertake a project similar to the current proposal, but would co-operate with us.

So far as I am aware, there is no copyright in the original sources, but consideration will have to be given to the issue of copyright in images produced from them.

LINKING IMAGES TO EXISTING TRANSCRIPTIONS

This is likely to be the first stage of the process in order to produce training material for the transcription process. An initial examination of the Selden resources indicates that there are a number (not yet established) of Latin transcriptions and associated translations which can be identified by Public Records Office references, for example CP40 = Common Pleas Plea Rolls, together with a Roll Index and membrane number. This will allow the Selden resource to be linked (after searching and visual inspection) to an AALT image.

A preliminary exercise might be to trawl all Selden resources and produce an inverted index of source references, if such an index does not already exist.

The primary Selden sources appear to be the English Reports (in 176 volumes). According to the Index Chart, these are grouped as follows :

Vols. 1 – 11 House of Lords (1694 to 1865)
Vols. 12 – 20 Privy Council (1809 to 1865)
Vols. 21 – 47 Chancery (1557 to 1865) **
Vols. 48 – 55 Rolls Court (1829 to 1865)
Vols. 56 – 71 Vice-Chancellors' Courts (1815 – 1865)
Vols. 72 – 122 King's Bench (1378 to 1865) **
Vols. 123 – 144 Common Pleas (1486 to 1865) **
Vols. 145 – 160 Exchequer (1220 to 1865) **
Vols. 161 – 167 Ecclesiastical (1752 to 1857), Admiralty (1776 to 1840) and Probate and Divorce (1858 to 1865)
Vols. 168 – 169 Crown Cases (1743 to 1865)
Vols. 170 – 176 Nisi Prius (1688 to 1867)

Some of the volumes marked ** appear to correlate with AALT images.

We will need to investigate if the above time-consuming procedures can be improved by automation.

Sample images are shown in the Appendix.

OUTLINE OF THE TRANSCRIPTION PROCESS

There already exists an AI-based transcription program – Transkribus (transkribus.eu), which would be free to use. There are a number of potential drawbacks in using Transkribus.

Firstly, the project would be reliant on external software which is not under its control. It is understood that part of Transkribus is open source, so that it could be moved to a local platform. However, part of Transkribus is proprietary software and without free access to it, this project is totally at risk. Transkribus is an EEC project, and if, following Brexit, access were either completely withdrawn or made chargeable to non-EEC countries, the only alternative solution would be to commission software to emulate those parts of the Transkribus package that had become unavailable. Secondly, Transkribus has a manual front-end which has to be used to feed images into the transcription engine. Besides (as a personal opinion) being somewhat user-unfriendly, it will not be practical when trying to process many millions of images. What is needed is to be able to submit a list of online locations of images directly to the engine which will then pre-process and process those images automatically. It must therefore be decided whether to produce an AI transcription program specifically for use by the project.

The pre-process involves identifying the document itself and extracting it from any irrelevant parts of the image, such as frames, supports and the like. It is not necessarily guaranteed that the document will be geometrically parallel to the image. (An example is the extraction of a number-plate from an off-centre image of the remainder of the vehicle.)

The next stage is to identify and separate the hand-writing from the background. It is assumed (at this preliminary stage, at least) that a monochrome image will contain all the information and that ink or membrane colour is irrelevant. Marks, tears, crossing-out, underlining and the like can complicate the process. A random sampling from the AALT website suggests that in many cases, there is damage and/or badly faded ink. Next, lines and individual words, including abbreviations, have to be identified, assuming that this will be more efficient than attempting to identify individual characters. Ascenders and descenders may interrupt clear line separation. An extra complication is to identify marginal additions.

Proof-of-concept experiments using Transkribus have suggested that, at this point in time, the Transkribus system is not effective enough in separating the text from the background.

At the end of the pre-process stage, it may be necessary for the program to refer back to a human editor to resolve elements on the page which it has not been able to pre-process with sufficient confidence.

The main transcription process stage will now involve using deep-learning to match the elements of the text into meaningful language. If the text is to be translated into English, it has to be decided at which point abbreviations will be expanded. The program should ensure that the text is grammatically and syntactically valid. The transcribed text will then be returned to a human editor for checking and correction. It is assumed that each element in the text will be assigned a confidence tag, which might be expressed visually as differently coloured text to assist the editor. After correction, the text is returned to the computer so that its deep-learning can be improved.

It is envisaged that the transcription process may not necessarily be serial, but rather more in batch format. A batch of texts will be transcribed and then checked for accuracy. Maybe the editing will be carried out by a number of editors who are not directly controlled by the project, but their work will be supervised by project moderators before being accepted. This could be an ongoing cyclical procedure with all “open” texts being continuously being improved in real time.

The point should also be made that texts will have be written by many different writers (“clerks”) and their hands may differ both vertically and horizontally, that is to say, the program may be able to benefit from indentifying the same clerk over a span of time (vertical) and also by applying its recognition of similar hand-writing styles of different clerks at the same period in time (horizontal).

OUTLINE OF THE TRANSLATION PROCESS

After the text has been satisfactorily transcribed, and it is not in English, it can then be translated. This process is well-established by online AI translation programs such as Google Translate. Google Translate already provides for Latin to English, but it may be necessary to adapt that to mediaeval legal Latin. Alternatively the project may find it more effective to produce its own translator, bearing in mind the relatively limited vocabulary of law court reports and legal transactions. If not already done, abbreviations (mentioned above) will need to be expanded.

As with the transcription process, it is envisaged that editors and moderators will check and enhance translations (as before, tagged with confidence markers) and feed any improvements back into the translation engine.

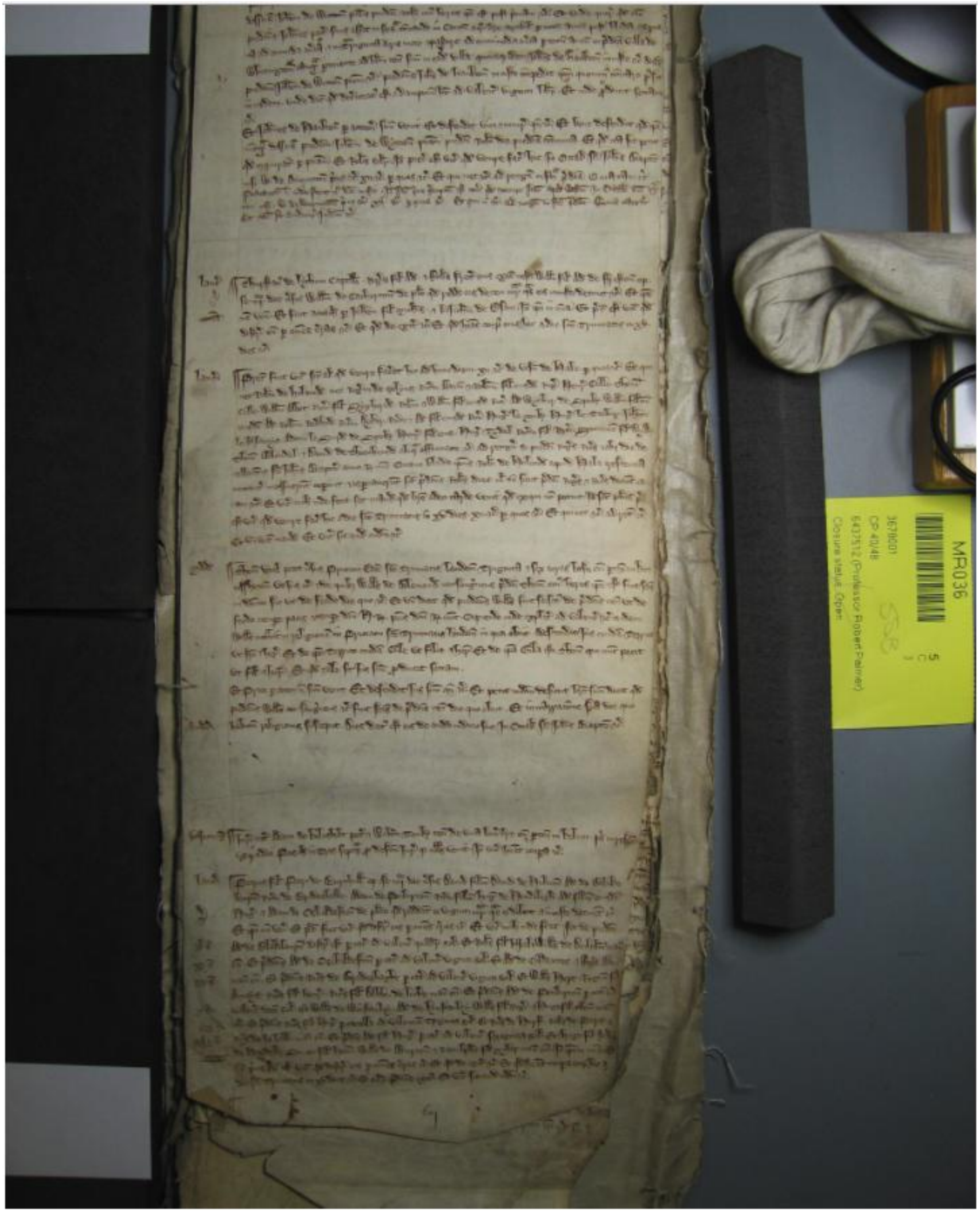
ANNOTATING AND ANALYSING THE ONLINE TEXTS

The final stage of the project will be to annotate each text so that there is an audit trail through the image back to the original source, so that researchers can verify for themselves the transcription and translation. Further analysis might also be useful. Obviously this will include the date of the original text together with the court or other authority generating the text. Where the text refers to a place or person, these can also be added. A comprehensive and flexible search engine should allow multiple ways of interrogating the body of texts so that not only individual words and phrases can be found but also patterns can be detected and analysed.

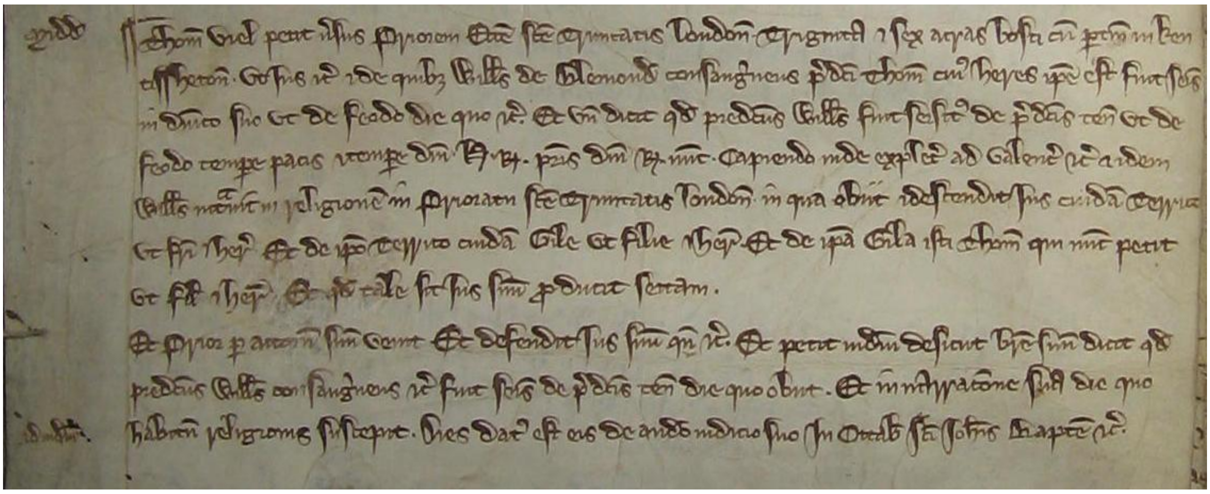
It may be necessary to specify conditions of use for the end product.

Ian Trackman : (ian.trackman@kenwoodvolunteers.org.uk)
13 June 2018

APPENDIX – sample texts


MR036
5
C
367601
CP 40/48
643512 (Professor Robert Palmer)
Closure status: Open

AALT image CP40no48/CP40no48afr/IMG_6976



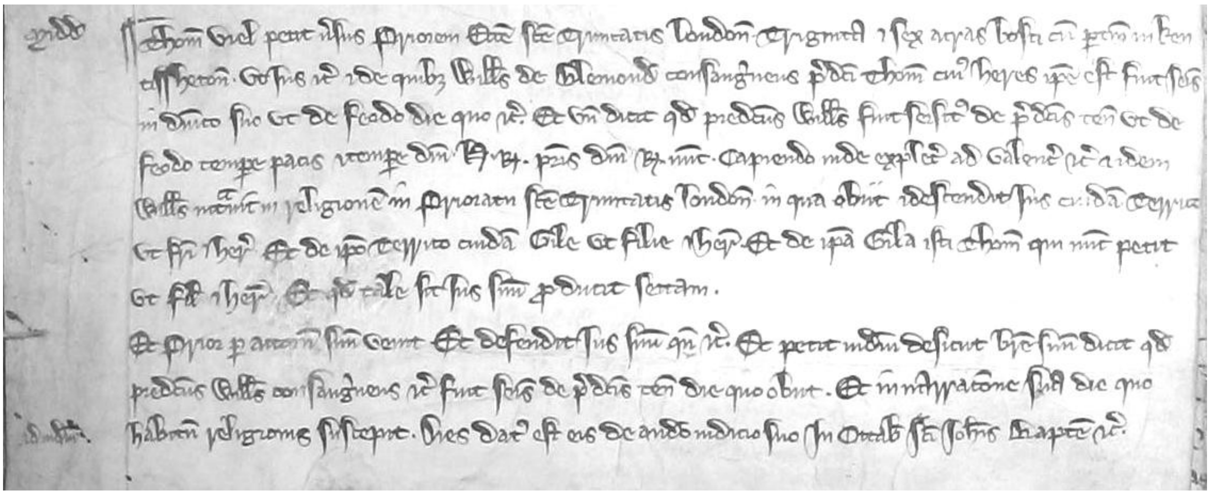
Middx

Thom' Gret petit v'sus Priorem Eccl'e s'ce Trinitatis London' messuagium j & sex acras bosci cu' p'tin' in Ken
tisshetone. Ut ius &c. de quibz Will's de Blemond consanguineus p'd'ci Thom' cuius heres ipse est fuit seis'
in d'nico suo ut de feodo die quo obiit. Et unde dicit q'd predictus Will's fuit seisit' de p'd'cis ten' ut de
feodo tempore pacis tempore d'ni H. r' patris d'ni R' nunc. capiendo inde explet' ad valenc' &c. Et idem
Will's intravit in religionem in Prioratum s'ce Trinitatis London' in qua obiit. Et descendit ius cuidam Terric'
ut fil' & her'. Et de ipso Terrico cuidam Gile ut filie & her'. Et de ipsa Gila isti Thom' qui nunc petit
ut fil' & her'. Et q'd tale sit ius suum p'fert sectam.

Et Prior p' attorn' suum venit. Et defendit ius suum qu' &c. Et petit iudicium desicut ip'e sum' dicit q'd
predictus Will's consanguineus &c. fuit seis' de p'd'cis ten' die quo obiit. Et in narratione sua dic' quo

[illegible] habitum religionis suscepit. Dies dat' est eis de audiendo iudicio suo In Octab' s'ci Johis Bapt'e &c.

Section extracted manually from above



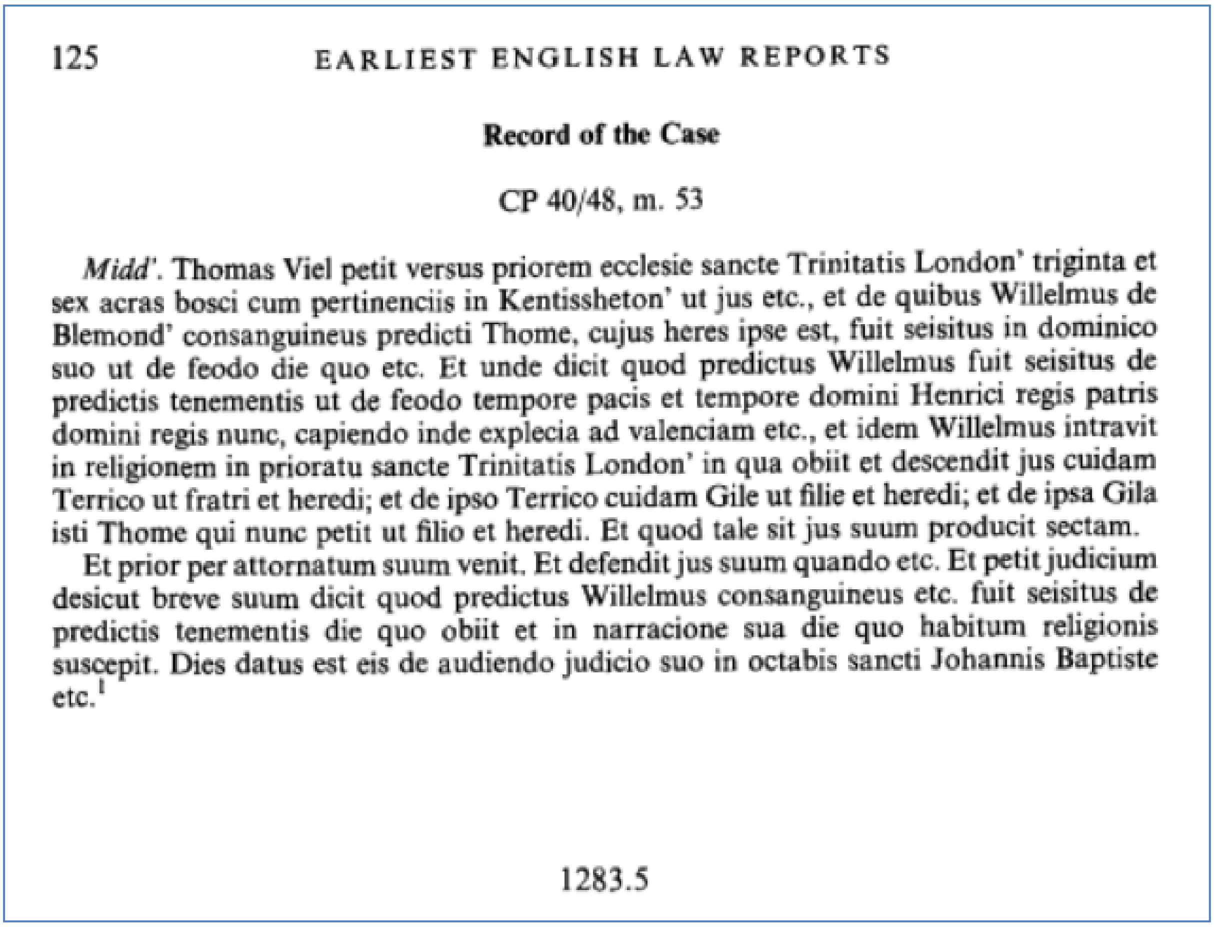
Middx

Thom' Gret petit v'sus Priorem Eccl'e s'ce Trinitatis London' messuagium j & sex acras bosci cu' p'tin' in Ken
tisshetone. Ut ius &c. de quibz Will's de Blemond consanguineus p'd'ci Thom' cuius heres ipse est fuit seis'
in d'nico suo ut de feodo die quo obiit. Et unde dicit q'd predictus Will's fuit seisit' de p'd'cis ten' ut de
feodo tempore pacis tempore d'ni H. r' patris d'ni R' nunc. capiendo inde explet' ad valenc' &c. Et idem
Will's intravit in religionem in Prioratum s'ce Trinitatis London' in qua obiit. Et descendit ius cuidam Terric'
ut fil' & her'. Et de ipso Terrico cuidam Gile ut filie & her'. Et de ipsa Gila isti Thom' qui nunc petit
ut fil' & her'. Et q'd tale sit ius suum p'fert sectam.

Et Prior p' attorn' suum venit. Et defendit ius suum qu' &c. Et petit iudicium desicut ip'e sum' dicit q'd
predictus Will's consanguineus &c. fuit seis' de p'd'cis ten' die quo obiit. Et in narratione sua dic' quo

[illegible] habitum religionis suscepit. Dies dat' est eis de audiendo iudicio suo In Octab' s'ci Johis Bapt'e &c.

Manually adjusted contrast of monochrome image



125 EARLIEST ENGLISH LAW REPORTS

Record of the Case

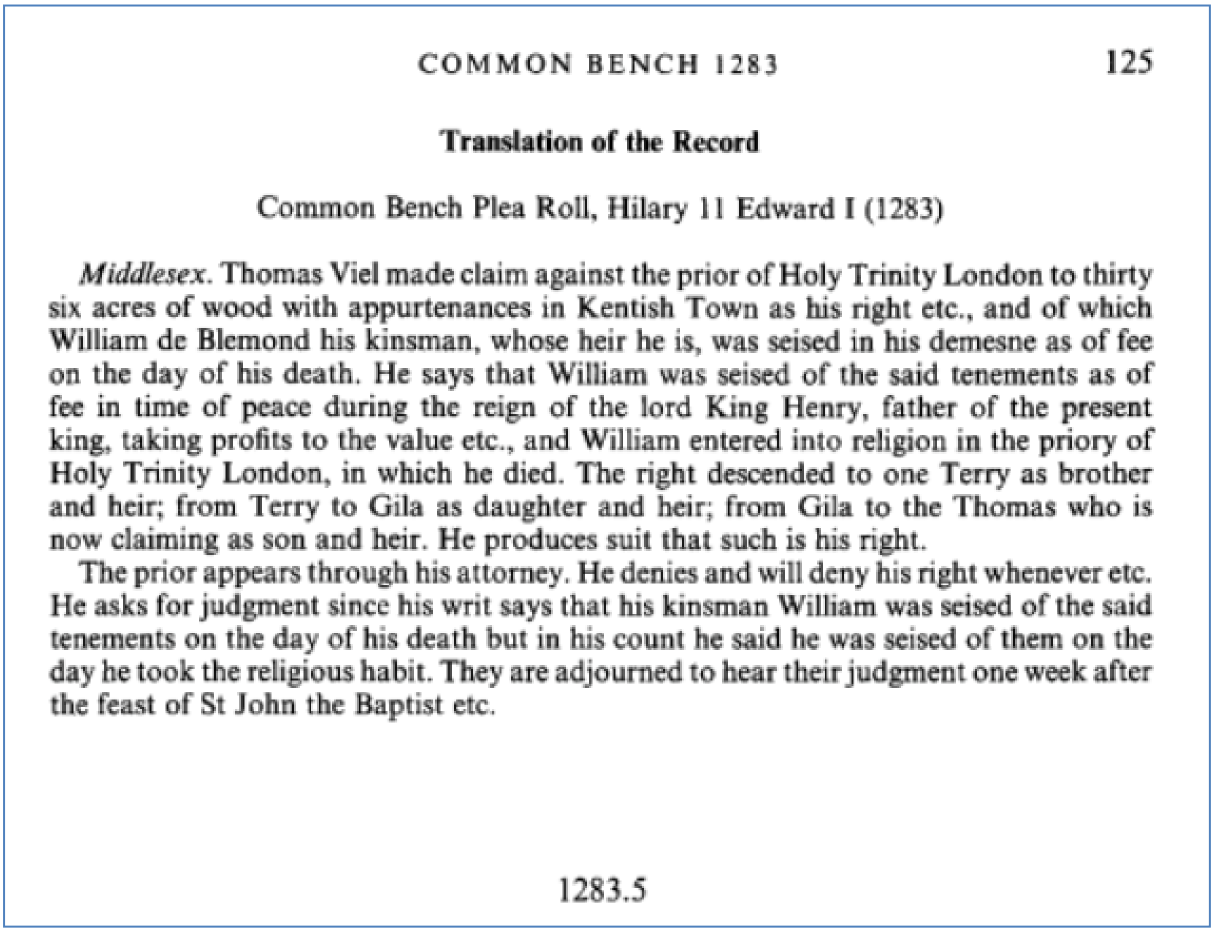
CP 40/48, m. 53

Midd'. Thomas Viel petit versus priorem ecclesie sancte Trinitatis London' triginta et sex acras bosci cum pertinenciis in Kentissheton' ut jus etc., et de quibus Willelmus de Blemond' consanguineus predicti Thome, cujus heres ipse est, fuit seisitus in dominico suo ut de feodo die quo etc. Et unde dicit quod predictus Willelmus fuit seisitus de predictis tenementis ut de feodo tempore pacis et tempore domini Henrici regis patris domini regis nunc, capiendo inde explecia ad valenciam etc., et idem Willelmus intravit in religionem in prioratu sancte Trinitatis London' in qua obiit et descendit jus cuidam Terrico ut fratri et heredi; et de ipso Terrico cuidam Gile ut filie et heredi; et de ipsa Gila isti Thome qui nunc petit ut filio et heredi. Et quod tale sit jus suum producit sectam.

Et prior per attornatum suum venit. Et defendit jus suum quando etc. Et petit judicium desicut breve suum dicit quod predictus Willelmus consanguineus etc. fuit seisitus de predictis tenementis die quo obiit et in narracione sua die quo habitum religionis suscepit. Dies datus est eis de audiendo judicio suo in octabis sancti Johannis Baptiste etc.[1]

1283.5

Selden transcription of the original



COMMON BENCH 1283 125

Translation of the Record

Common Bench Plea Roll, Hilary 11 Edward I (1283)

Middlesex. Thomas Viel made claim against the prior of Holy Trinity London to thirty six acres of wood with appurtenances in Kentish Town as his right etc., and of which William de Blemond his kinsman, whose heir he is, was seised in his demesne as of fee on the day of his death. He says that William was seised of the said tenements as of fee in time of peace during the reign of the lord King Henry, father of the present king, taking profits to the value etc., and William entered into religion in the priory of Holy Trinity London, in which he died. The right descended to one Terry as brother and heir; from Terry to Gila as daughter and heir; from Gila to the Thomas who is now claiming as son and heir. He produces suit that such is his right.

The prior appears through his attorney. He denies and will deny his right whenever etc. He asks for judgment since his writ says that his kinsman William was seised of the said tenements on the day of his death but in his count he said he was seised of them on the day he took the religious habit. They are adjourned to hear their judgment one week after the feast of St John the Baptist etc.

1283.5

Selden translation of the original