A Digital Edition of a Spanish 18th Century Account Book:
User Driven Digitisation

Abstract

In this, part one of a two part paper, we will discuss the approach taken during the creation of a digital edition of the *Alcalá Account Book* manuscript. The *Alcalá Project* was originally proposed as a digital humanities project to mark a humanities collaboration between the *University of Alcalá de Henares* (UAH), Spain, and the *National University of Ireland, Maynooth* (NUIM). The source was to be a Spanish eighteenth century account book recording the monthly expenses of the Royal Irish College of Saint George the Martyr. In the given time-frame, the source manuscript was chosen, encoded and made available in a web based, dual language, searchable and interactive environment. More importantly, a virtual framework was developed to aid the historian in answering historically pertinent research questions that are specifically prompted by the historical object – an account book.

We contend that by creating a digital edition that represents the original *Alcalá Account Book* manuscript and its functionality, we have provided the end-user (for example the historian) with a richer environment for performing research, and a research tool that is specifically designed to be fit for purpose. The approach is informed by the discipline of the participants, and also takes account of current best practice in humanities and digital humanities. We aim here to provide a description of that digitisation process and the methodologies we used to make design decisions. Part two, *Formalisation and Encoding*, will address the theoretical framework and practicalities of formalising and encoding the source.

[1] **Historical Background and Significance**

[2] The *Royal Irish College of Saint George the Martyr* (*El real colegio de San Jorge Mártir de los irlandeses*) in Alcalá was founded in 1649, with ten students and a small staff that included a rector, two non-resident professors and two or three servants. Its primary function was the education of Irish students for the priesthood. It was closed down by order of King Carlos III in 1785 as part of a rationalisation plan and as a result of the decline in both student numbers and discipline. It was amalgamated with the Irish college in Salamanca, which itself closed in the middle of the last century. Since then the college buildings and grounds have become the property of the *University of Alcalá* (which was itself reconstituted in 1977, following its suspension in the early nineteenth century). The University is currently home to 25,000 students and nearly 2,000 academic staff.

[3] The digitised material presented here is taken from the college's account books or *Libros de gastos del colegio de Alcalá*. [1] They were placed in the archives of the Irish college at Salamanca on the closure of the *Alcalá college* in 1785 and were brought to Ireland in 1951. They are now housed in the *Russell Library*, *Maynooth College*, which houses printed books, manuscripts and archives, as well as a conservation department. The account books now form part of their Salamanca Archive.

[4] The accounts, which cover the period 1774 to 1781, offer a unique insight into the day-to-day running of the College with valuable information on diet, discipline and domestic matters. They consist of 324 pages, structured in 63 folios, that is, sheets of paper folded to make several leaves in the codex. They record the ordinary and extraordinary expenditure at the college each month, and also detail information on the number of students and servants provided for. Five of the folios provide summaries of previous monthly records and were prepared for the auditor of the books.

[5] **The Digital Edition**

[6] The online digital edition, available at [1], provides a high-resolution facsimile, Spanish transcription and English translation for each of the 324 pages. These are presented in an interactive and dual language environment along with data-sheet functionality to support accounting operations. Supporting documentation is also provided. It adheres to all six of Vanhoutte's criteria for the creation of a digital edition. [2]

[7] The potential audience of historians for this type of document ranges from palaeographers, through religious scholars, to social historians. A participatory approach, where end-users are actively involved in the design process, was adopted for the development cycle, which led us to prioritise the provision of functionality for social historians researching the Early Modern period. [3] Our end-users are supported by providing for the selection, retrieval and storage of, and subsequent computation on, the expenses contained in the Account Book. However, detailed palaeographic work can also be undertaken using the facsimile images in conjunction with ›zoom‹ and ›pan‹ functions.

[8] The *Alcalá Account Book* details the expenditure over a seven-year period from 1774 to 1781 (though the accounts for 1780 are missing). Ordinary and extraordinary expenditure is detailed separately; typical categories include bread, wine, meat, the salaries of the laundry-woman and cook, and petty expenses. Browsing the digital edition for items of interest is possible, given that it lends itself to random access, not just serial access. [4] In addition, there are focused searching tools provided, for example, specific items can be located using a keyword search that acts as a filter on all of the pages of the Account Book. It performs a Boolean search operation on the space separated terms in the search box, then presents the resultant thumbnail images in a panel for selection and further manipulation. [5] Regardless of whether the Spanish or English interface is in operation, it is always possible to search the Spanish transcription or English translation. A selected page is presented as full facsimile, transcription and translation where the sought keywords are highlighted, depending on which language was used for the search, this is illustrated in Figure 1.

[9]

| Results:89 | Search: bread wine | in English ▼ | Search | Language: English ▼ | To Datasheet |

Image View | **Spanish Transcription** | **English Translation**

Ordinary expenses for the Month of December 1774 in which there were 14 Scholars and two Servants who at the ratio of a Pound of **Bread** another pound of Meat and a Cuartillo mayor of **Wine** for each of the aforesaid amount to the following items and sums

	Description	
☑	**Bread** for same are 252 and a half at 9 cuartos amount to	0267 12
☐	Meat for same is 352 Pounds at 11 cuartos amount to	455 18
☐	**Wine** for same is 15 arrobas and a half 9 and a half at 14 Reales and a half and 4 maravedes and the remainder at 12 Reales and 4 maravedes amounts to	0211 24
☐	Abstinence days are 144 at 12 cuartos	0203 09
☐	Spices amount to	0019 10
☐	Two reales a day for the Rector	0062 00
☐	Salaries of Cook and Laundrywoman	0035 00
☐	Two arrobas 11 Pounds and a half of cured Bacon at 56 Reales an arroba and 4 Pounds of Fresh Bacon at 14 cuartos	0144 12
		1396 17
		1358 19

Extraordinary Expenses for said month

	Description	
☐	One Pound of Wax	0010 00

Thumbnails: f013-02, f014-02, f015-02

[10] Figure 1. »Bread« and »wine« used as keyword filters, items of interest selected in English translation on one resultant page.

If the encoding were viewed in a generic viewer, for example, a web browser displaying the image and XSLT-processed text, the user would be able only to read the rendered text. If viewed using our custom-developed software, a Flash application designed to support the user in researching the *Alcalá Account Book*, there are numerous additional features of the document presented, and requirements supported. An example of this is the Select function, which can be used to transfer expenses to the data-sheet. This is operated with a check-box in the text-versions, or by clicking on the relevant entry on the facsimile. There are up to five data-sheets available for use. These allow for the sum, average, maximum and minimum of the list of financial entries to be calculated, thus supporting the user in their research.

[11] **Digitisation Process**

[12] **Image Capture and Management**

[13] The digitisation process commenced with image capture; 324 pages, roughly 21cm x 30cm in size, were imaged in the *Russell Library*, each one stored in TIFF format. Apart from the significance of their dates, which cover the final years before the college in Alcalá closed, they were also more suitable for imaging than other folios as they were unbound and so were at less risk of damage.

[14] Despite illuminating the manuscript with fluorescent cold lighting the brightness of the images varied because of changing ambient light within the library. The images therefore required post-processing to provide uniform brightness across the pages and collection. This post-processing, together with centering and cropping, was accomplished using standard image manipulation software. [6] The preservation quality images were then used as the source for a repository of JPEG image tiles to be used in the eventual image viewer. One function of the viewer is to dynamically render an interactive facsimile of each page at various zoom levels. This required the generation of 168,219 image tiles, i.e. 324 pages having 5 zoom levels, and on average 103 tiles per original image.

[15] A selection of images was chosen to prototype the transcription, translation, segmentation and encoding processes. Rapid prototyping of these processes on this small sample allowed for each process to inform the performance and refinement of the others; as the segmentation of the image was performed it helped to clarify the structure of the document, which was, itself, only fully understandable after some translation had been provided. Once we ›were confident that these processes were reliable each one could be applied to the whole of the document.

[16] **Transcription and Translation**

[17] So as not to restrict the usage to Spanish-speaking researchers, a translation of the *Alcalá Account Book* was also required. The Spanish in the Account Book was archaic, dating from the eighteenth century and there were also esoteric terms that resulted from its religious context. In order to achieve a high-quality transcription and translation it was essential that the translator possessed appropriate linguistic, palaeographic and historical skills. Although these expertise were available within the research team, independent validation was also sought from external consultants.

[18] The goal was to create a diplomatic transcription of the Spanish text. The transcription preserved inconsistencies in spelling, abbreviation, capitalisation and punctuation. Consequently there are implications for usability, for example a keyword search for »dichos« will not return the page that contains the abbreviated version, »dhs«. Of course, there are ways to overcome this, for instance, normalisation could be performed within the encoding. This was not considered sufficiently pressing to warrant attention in the first full iteration of design.

[19] The translation mirrored the transcription in cases where meaning was not obscured. Some normalisation was performed; abbreviations were expanded and direct translation was eschewed if clarity was compromised, for instance »Bread for 305 scholars«, followed by, »Said's meat«, has been changed to, »Bread for 305 scholars«, followed by, »Meat for same«. Furthermore, given that translation was involved there could be different results for a search on a particular Spanish word and its direct translation e.g. »pan« does not translate as »bread« in the case where »pan« forms part of another Spanish word, so a search on »bread« might not produce the same number of results as a search on »pan«.

[20] Translation proved to be difficult for a variety of reasons. The scribes occasionally changed vowels and consonants. for example »lave« instead of »llave«, which resulted in different pronunciation. Given the absence of a wider context, the interpretation was made difficult. Polysemy was another problem, for instance in Spanish »fuente« can refer both to »fountain« and »large dish«. Lastly, the historical and religious context described above meant there were some obscure words, phrases and abbreviations in the original. One example that remains untranslated is »Emmdo – c – d – seis – end – Vale«, which appears to be an official abbreviation or correction on the notary copies.

[21] The decision was undertaken to indicate any uncertainty deriving from difficulties in translation using a visual cue in the English text version. Furthermore any uncertainty derived from an inability to accurately ascertain the Spanish word or phrase was indicated in the Spanish text version, though this is very uncommon. There are approximately 38,000 words in the translation and approximately 30 instances of questionable translation. There are approximately 39,000 words in the Spanish text-version, and approximately 10 instances of questionable transcription.

[22] **Encoding**

[23] Based on the prototyped sample, and using those transcriptions and translations it provided, it was possible to establish an encoding process. This consisted of segmenting the *Alcalá Account Book* (identifying semi-autonomous sections of the accounts, for instance, paragraphs, expenses, etc), image-mapping the segments (recording the coordinates), and structuring and encoding the Extensible Mark-up Language (XML). [7] As we moved forward in the project it was possible to supply the translator with the segmented images. The returned transcribed and translated products were encoded using the developed XML Schema. [8]

[24] Our approach is to model the structure and meaning of the work instantiated by the document (the logical model), this is then placed in its physical context to create another model, and this model is eventually placed in its digital context to create the digital edition. This means that our first model is designed around what McGann and Buzzetti call »the work«, [9] which is known as the *Alcalá Account Book*, rather than the underlying the instance of it captured by the folios of manuscript. The *Alcalá Account Book* is a refinement of the abstract class that defines an account book. This allows us to use a heuristic model of an account book to inform an initial modeling and segmentation of the work. This segmentation then acts as a context for general inspection of the document to identify model elements, for example, an expense item, a signature, a heading or a monetary amount. A material segmentation is then performed, replacing heuristic elements with actual components, and thus informs the refinement of the heuristic model. A hermeneutic spiral is created so that subsequent iterations of model-informed segmentation act to refine the model and segmentation process. [10] Ultimately, a contextually appropriate model and correlated segmentation approach are produced. This process, occurring simultaneously in the material and the mental realms, is illustrated in figure 2.

[25] Once the model for segmentation has been developed to the point where it can be structurally represented as a tree, it can be translated into any schema or language. [11] This is an account book tree, not a document tree: segmentation of individual pages involves the association of textual elements on the physical page with nodes on the account book tree. In practical terms, this required the material identification of segments on the captured image using software to create polygons, the coordinates of which were recorded for later use in the application.

[26]
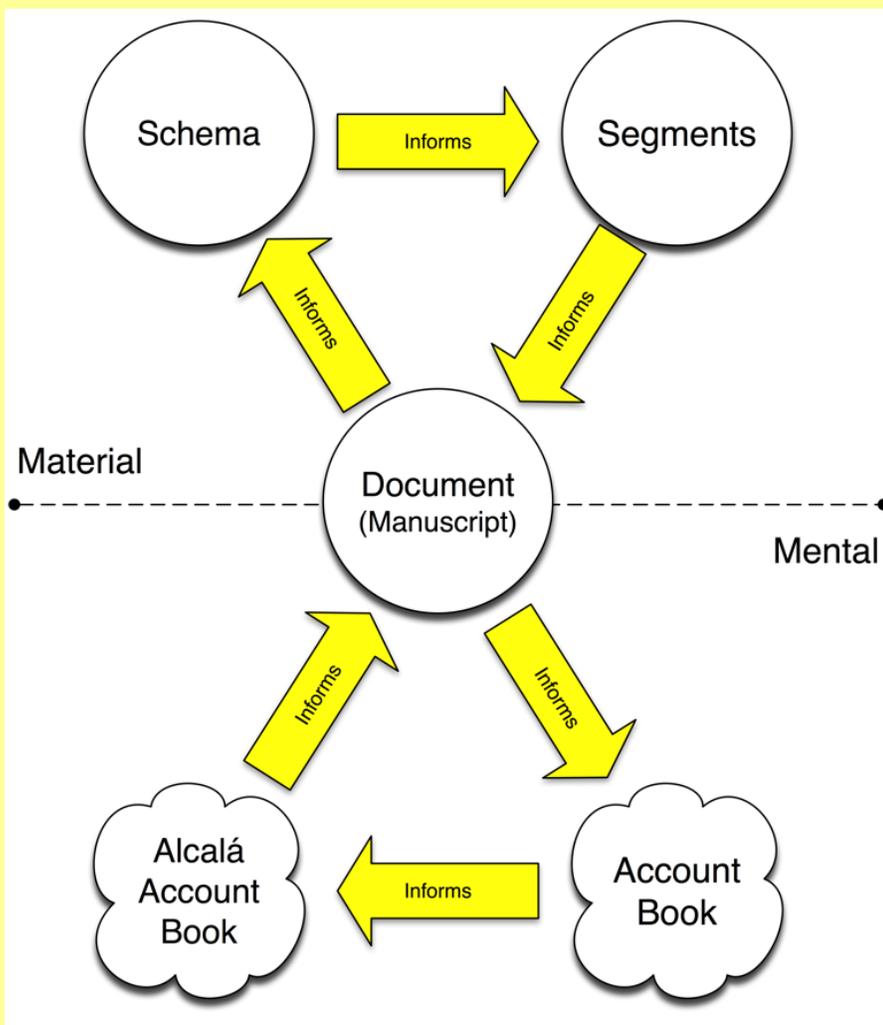
[27]   Figure 2: Hermeneutic spiral of understanding and interpretation is used to create segmented model (and schema) of the *Alcalá Account Book* manuscript.

[28] This approach was, in the main, successful, but relies on the choice of exemplars used in the sample for prototyping. In practice, there were occurrences of items in the *Alcalá Account Book* manuscript that did not appear in the initial sample, for example, quarterly summaries of previous accounts. However, it was possible to refine the model to include these summaries as part of the encoding process.

[29] The production of this resource stems from a software engineering approach to resource development. It considers the encoding and the software in conjunction with each other, neither in isolation. This is an approach, rather than a solution; it is one example of how to use these software engineering practices in an encoding project.

[30] These are the main characteristics of our user-driven encoding: that the work, rather than the document, formed the basis of the logical model, and that the functionality embodied in that work is supported by the encoding. Further, if the Use Cases require interaction with the physical model (the manuscript) this should also be supported. [12] For instance, we wish to be able to present the corresponding facsimile image for a month of particular interest, so this requires that the software can extract the correct image reference from the encoding. In turn, our user-driven digital edition has two characteristics: that it is based on a user-driven encoding, and that it provides the functionality embodied in the *Alcalá Account Book* manuscript, along with any additional user requirements related to the digital edition, for instance, dual language capability. It was necessary, therefore, to identify an encoding language that was capable of supporting a user-driven encoding but could further support a user-driven software environment.

[31] Our encoding, the rules of which were to be expressed in a schema, had to support all three aspects of the user-driven digital edition: the segmented model of the *Alcalá Account Book*, the manuscript, and the original plus digital user requirements fulfilled by the software environment. The user-driven logical model of the work, the *Alcalá Account Book*, acts as a mediating artefact in the encoding. [13] It represents the logical structure of our source, and it can be linked to the physical structure of the manuscript and the user-requirements fulfilled by the interactive software, but also embodied in the logical and physical models. It is the foundation class for building the other features and methods of the physical and interaction classes. It was only by designing a custom schema that we were able to create this mediating artefact. As a consequence, we chose XML as our encoding language because it supports custom schema.

[32] Deciding to create a user-driven digital edition means that we are supporting the functionality embodied the *Alcalá Account Book*, rather than just the viewing of the manuscript. In addition, we are supporting user-requirements specific to the digital edition, such as dual language provision, searchability and manipulation of datasheets. The encoding expressly supports the software application by associating the transcription, translation and image-map co-ordinates of each segmented textual element. This allows for the correct rendering of the text in Spanish and English as the various element names indicate to the software how they should be rendered on screen. It also supports the select function on the facsimile and text versions, allowing various elements of interest to be transferred to the data-sheet. These aid the users in their research and thus meet our objectives.

[33] The translations are provided as a support to the digital edition, to the extent that we do not have two separate files for the transcription and translation encodings. Instead, each text segment is represented in its transcription and translation, along with its coordinates. Text segments that do not require translation exist only as transcriptions, for instance, monetary amounts and signatures. This form of segmentation and encoding allows for the software to switch easily between Spanish and English representations and thus supports the dual language needs of our users.

[34] We were able to create user-driven digital edition by choosing element names for our encoding that embodied a description of, and uses for, the textual elements on the page. Meeting current user-requirements is one benefit of this approach; another is the flexibility the approach provides. After a user trial it was requested that in addition to the expense items themselves, the total expenditure for each month should be selectable using the checkbox or mouse-click. Given that each segmented element was custom named using an XML element this did not require any change to the encoding and only a small change in the supporting software.

[35] ## Information Architecture and Visualisation

[36] The goal of the information architecture design process was to store a digitised representation of the *Alcalá Account Book* manuscript in an extensible online digital repository and to provide a client application that could present an interactive version of this stored digital object to the user. Both the repository and client were implemented to be flexible; for example the rendering and the interactive user interface dynamically generated by the client were derived from the XML encoding. Furthermore, the *Alcalá Account Book* and the abstract class of an Account Book were used in conjunction with the XML to dynamically construct an interactive version of the *Alcalá Account Book* manuscript using the client. The significance of this approach is that future data models of further material from the Salamanca archive can be incorporated into the architecture with minimal program code redevelopment for the client software.

[37] The digital version of the *Alcalá Account Book* manuscript is a collection of three separate data sets: the XML encoding of the 324 pages of the *Alcalá Account Book* and its manuscript, the 168,219 images required to dynamically render an interactive facsimile of each page and lastly, the metadata of this digital version. As part of this project, a series of information storage and retrieval experiments was conducted and it was determined that a single software application would be inefficient in handling both the complex XML querying and high volume image storage and retrieval for this application. It was concluded that these three data sets should be managed using separate software, each optimised for that particular data set. Finally, the publicly accessible interface is provided by one entity thereby presenting a single interface to the user.

[38] We examined a number of popular digital repository solutions that did not meet our requirements. *Eprints* [11], a popular digital repository solution, does not provide support for fetching data from external software, and by design it is configured to focus solely on hosting academic output for example preprints, postprints and theses. *DSpace* [12], unlike *Eprints*, can host a variety of user definable digital objects and collections but does not have the capabilities to act as a manager of digital objects whose elements are stored in separate external servers. *Fedora* (Flexible Extensible Digital Object Repository Architecture) [13] was chosen as the managing entity of our information architecture as it could host any custom-designed digital object and it can provide additional services through a process known as dissemination. Services may be added to a digital object in *Fedora* that point to data and queries stored in other storage solutions.

[39] In addition to requirements outlined above, there were additional benefits to be had from choosing *Fedora*. Firstly, it is mature and stable open source software that has been used extensively for online historical archives such as the *Encyclopaedia of Chicago* [14] and *University College Dublin's Irish Virtual Research Library and Archive* [15]. Secondly, the platform supports a large number of database cluster configurations, necessary for future scalability. Finally, it also provides an interface implementing the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) [16], which is provided by the internal indexing facilities of the *Dublin Core* [17] format. *Fedora* also supports ingestion using FOXML [14] or METS [15]. The *Fedora* digital objects for this project contain the high-quality captured images together with text encoding and associated metadata, all of which are harvestable using OAI-PMH.

[40] While it is possible to store XML directly within the digital objects, *Fedora* does not support advanced querying capabilities within the XML itself. Therefore, an XML database was included in the architecture to store the XML encoding of the *Alcalá Account Book* manuscript and to provide XQuery functionality for data retrieval from the encoding. *Fedora's* external communication via disseminators must be implemented using the Hypertext Transport Protocol (HTTP). [16] High quality commercial solutions such as Oracle did meet our technical requirements but an open-source software was preferable due to our budget constraints. *eXist-db* [18] was chosen as it is stable and provides a full implementation of the XQuery 1.0 standard, optimised for fast data retrieval through a HTTP service. A performance-tuned *Apache* web server was deployed to host and deliver the image collection. *Fedora*, as manager, firewalls public access to the *eXist-db* and *Apache* servers. A combination of SOAP [17] and HTTP were used for information exchange with *Fedora*. Each digital object stored in *Fedora* has a unique identifier known as a Persistent Identifier (PID), which was also used in this project to identify the object's elements stored in *eXist-db* and *Apache*.

[41] As mentioned earlier, the functionality of the client software was dynamically generated from the structure of the XML document retrieved from the repository. Therefore, it was important for the client to identify the class of the digital object. With this architecture, class identification is possible using three distinct processes: by identifying the class from the metadata of the object, by deriving the class through automated inspection by the client of the XML document or by identifying the class from the services provided by the *Fedora* digital object. The set of disseminators, that is, the methods available to the class, were labelled at creation with the name of the class. The most efficient method, therefore, was to simply determine the class by querying the label of the set of disseminators. This meant that it was possible to determine the class of object without ever accessing the XML, reducing the overhead of querying the XML server and preventing the client from querying the XML until it had identified the queries associated with the digital object.

[42] The client software is composed of several modules, each charged with a specific function, as follows. The communication and parser modules are responsible for communicating with the repository, and supply the retrieved XML to three core modules: the logical, interaction and physical model modules. The modules are used for the dynamic development and delivery of the most appropriate interface, interaction mechanisms and graphical representation of the page under consideration. The representation and, ultimately, direct manipulation of the user interface, is entirely driven by the XML and the suite of functional methods provided by the interaction module. This adaptable client is a key component in the delivery of a satisfactory user experience when research is conducted with the digital edition of the *Alcalá Account Book* manuscript.

[43] The development approach was to use *Rapid Application Development* (RAD) to produce a lightweight, browser-based application. [18] Typically, there are three choices for this type of implementation: generating a dynamic XHTML website, implementing a Java applet, or producing a Flash application. Investigation of the use of Java interfaces in existing repositories indicated that several had experienced problems with their Java client implementations. Most notably, ARTstor [20] recently ceased development of their Java applet client and is in the process of replacing it with a dynamic HTML version to take advantage of significantly greater flexibility, greater system efficiencies and improved performance and speed. Dynamically generated website front-ends for repositories are common, for example the *New York Public Library Digital Gallery* [21]. For our interactive application it would be necessary to use *Asynchronous JavaScript And XML* (AJAX) technologies if this approach was adopted. However, this would produce what is essentially a group of server applications interacting with a series of browser-based JavaScripts making HTTP requests. Although such a venture was within our expertise, in our experience it would be more manageable and faster to create a single client. Furthermore, developing this type of website requires extensive resources to test and adapt the client for different environments. In our experience it is almost always necessary to repeat this process for future releases of web browser software.

[44] It was decided, therefore, to develop the client application as an *Adobe Flash* application compiled with *Adobe Flex Builder 3* [22]. Flex Builder is a *Rich Internet Application* (RIA) development platform that can produce a single self-contained application file that can be run on a wide variety of operating systems and browsers. Furthermore, *Adobe Flex* was designed to efficiently manipulate XML and provides a powerful interface to handle SOAP communications, utilised to communicate to *Fedora*. The language also supports localisation for creating language user interfaces. It also has the ability to generate a dynamic user interface from XML; in fact the source code of a Flex application is written entirely in XML.

[45] The high-resolution TIFF image facsimiles of each page were too large to be delivered on-line as a single image. Using a compressed file format such as JPEG reduced the file size significantly, but they remained too large to be rendered smoothly in the client application. Reducing the resolution of the images would have reduced the file size further but it was preferable to present the image in the highest-quality resolution to allow the user to have the best view possible of the facsimile within the application. A solution to this problem was to display a subsection of the image at a specified zoom level at any one time. Google uses such technology for rendering *Google Maps* [23], for example. This technology was utilised in the project by integrating the *Zoomify Flash [24]* component into the client application. *Zoomify* has been used with numerous online archives such as the *University of Maryland Digital Archives* [25] and the interactive Flash exhibits in the *Library of Congress Exhibitions* [26] site.

[46] A typical Use Case for this system is detailed in the following scenario. Upon start-up, the client sends, using the digital object's unique PID, two queries to *Fedora*: (1) determines the languages of the text in the XML encoding, which the client uses to construct the search interface, and (2) obtains a list of physical page identifiers from the XML encoding which the client then uses to generate a list of thumbnails and initiate a download of these thumbnails from the server. Clicking on a page thumbnail initiates two concurrent processes. Firstly, it makes the request to *Fedora* to fetch the XML representation of the physical page of the digitised work, for example the client requests page f003–01 from the *fetchPage()* disseminator provided by the *Alcalá Account Book* manuscript digital object stored in *Fedora*. This disseminator has been configured to forward this request to the *getPage()* query stored in *eXist-db* database. The database executes the query, returning the result to the *fetchPage()* disseminator. *Fedora* then forwards the result to the client application, which parses the XML, renders the transcription and translation texts and provides a suite of interactive tools. Secondly, the client software makes a series of requests to *Fedora* to retrieve the image tiles required to render the facsimile image of the selected page. *Fedora*, in turn, forwards the image requests to *Apache*, and the resulting images are returned to the client, which renders the image. Subsequent panning and zooming of the image will require the client to request further image tiles from *Fedora*. A block diagram of the overall architecture is shown in Figure 3.

[47]
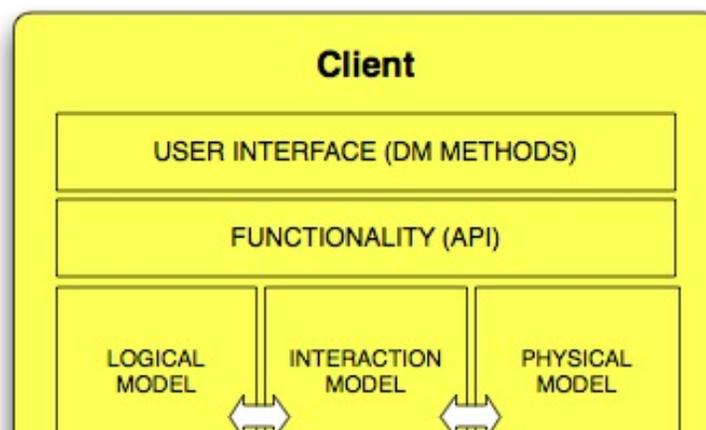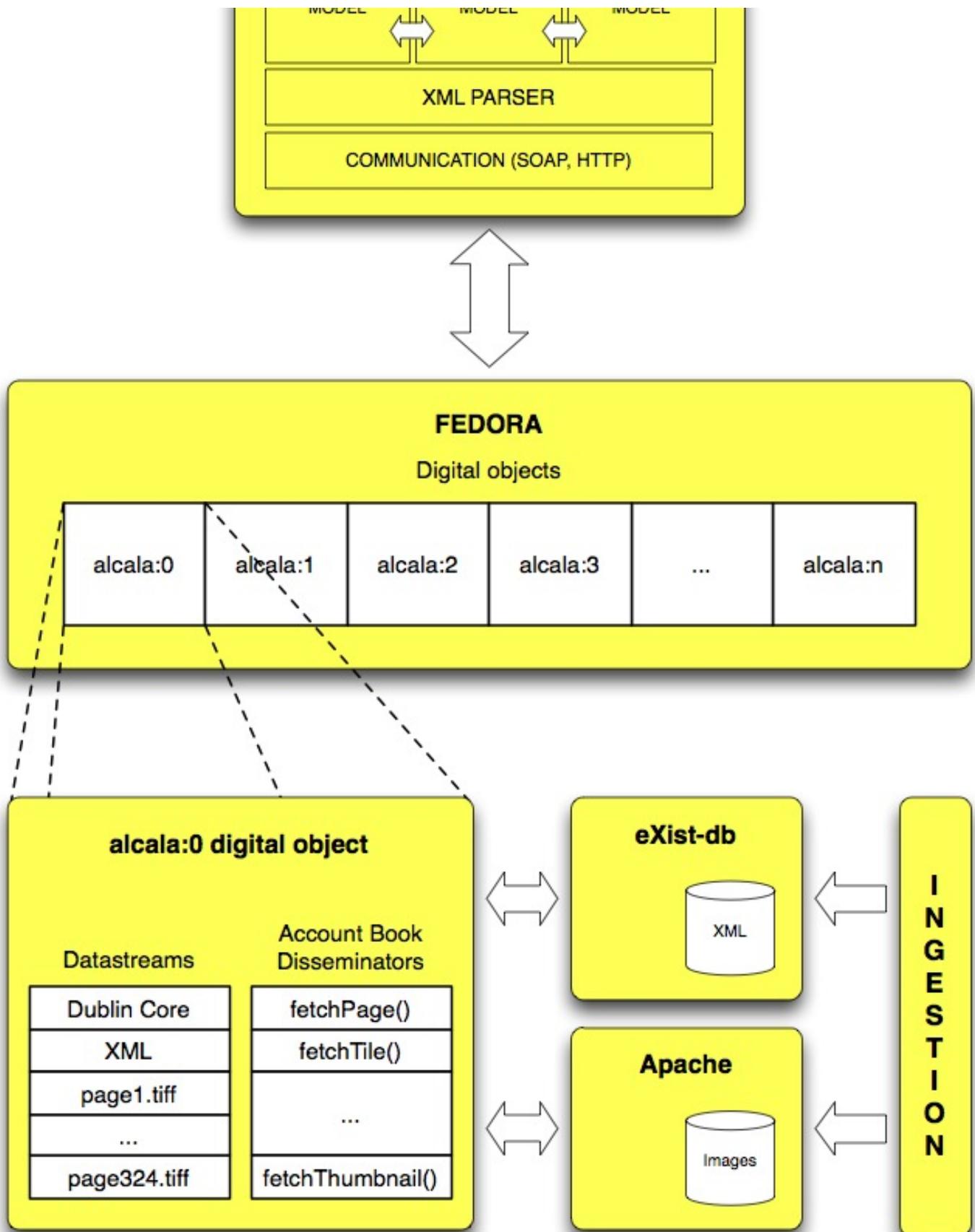
[48]   Figure 3. Block diagram of digital repository for the digital edition of the *Alcalá Account Book* manuscript

[49] **Conclusions**

[50] We argue that digitisation of historical artefacts should firstly preserve the usability of those artefacts, and then should add value to the artefact using the tools available in the digital world. We believe that this usability can only be provided by simultaneous consideration and design of the encoding and the accompanying software. In considering the encoding as an intrinsic part of the functionality delivered by the software, we elevate that encoding above its usual role of contextualisation. The encoding must emulate the internal rules, structures and meaning of the »work« being modelled first and foremost. It is then possible to place this »work« in its physical context, the document. It follows therefore, that the functionality (or usability) of the document, is preserved, making it truly accessible.

[51] The model that we have created and expressed in custom XML allows us to support the functionality of the original account book, as well as the functionality of the digital edition, for example, the simultaneous presentation of both the transcription and translation. The methodology that shapes the use of advanced software solutions is built upon the same methodology used to create the encoding and provides the functionality of the original work, physical document and digital edition.

[52] Our user driven approach is realised as a concurrent design process where we build the model, the encoding and the software using rapid prototyping, an iterative design process. Choosing XML as our encoding language meant that the software engineer spent less time trying to ›work around‹ the sequential approach favoured by TEI; adhering to a sequential design process often results in the biggest iteration of all, going back to the start. [19]

[53] Just as the encoding and software is flexible in order to meet the users' needs, so too is our information architecture. By using the Fedora Commons repository framework, we have ensured that each digital object, and every method required to interact with them, is modularised, and therefore easily updatable and extensible. We have also provided extensive metadata for harvesting so that our digital repository is ready for use by other repositories and software applications.

[54] We believe that our responsibility (as document encoders and software engineers) is to the interactive, rather than passive, user. For this project we had a specific purpose and audience; it was not our intention to create another machine-readable text. Our encoding could have been made viewable in a variety of contexts, without customised software, by the addition of an XSLT or CSS (like TEI or any other encoding system based on XML). However, in addition to rendering, we wish to provide functionality to support our users in actually using the source. In this way new research can be supported that is driven by the source, rather than a by-product of it. We hope that this may contribute to the ongoing debate surrounding the ›value of digitisation‹, for example, apart from preservation and presentation what new research has been achieved? In addition to providing a modern, flexible repository, we will promote and support active research in those sources that are selected for this type of user driven digitisation.

[55] **Acknowledgement**

John G Keating/Aja Teehan/Damien Gallagher/Thomas O'Connor
An Foras Feasa
National University of Ireland, Maynooth
Maynooth, Co. Kildare, Ireland
@

(17.06.2009)

Bibliography

Black, Paul E. (ed)

> 2004 Dictionary of Algorithms and Datastructures [online], U.S. National Institute of Standards and Technology [2].

Boole, George

> 1854 An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities. Macmillan. Reprinted with corrections, Dover Publications, New York, NY, 1958. [27]

Computer Professionals for Social Responsibility

> 2008 Participatory Design [3].

Fielding, R./J. Gettys/J. Mogul/H. Frystyk/L. Masinter/P. Leach/T. Berners-Lee

> 1999 Hypertext Transfer Protocol -- HTTP/1.1, Internet draft standard RFC 2616, June 1999.

Landa, José Ángel García

> 2004 Retroactive Thematization, Interaction, and Interpretation: The Hermeneutic Spiral from Schleiermacher to Goffman. In: BELL (Belgian English Language and Literature) ns 2: p. 155–66.

Laurillard, Diana

> 2002 Rethinking University Teaching – A Conversational Framework for the Effective Use of Learning Technologies, 2nd Edition. London and New York: Routledge, p. 86–89.

Martin, James

> 1991 Rapid Application Development. New York: Macmillan.

McGann, Jerome/Buzzetti, Dino

> 2006 Critical Editing in a Digital . In: Lou Burnard, Katherine O'Brien O'Keeffe, John Unsworth (eds): Electronic Textual Editing. New York: Modern Language Association of America, p. 53–73 [4].

Senko, M.E./Altman, E.G./Astrahan, M.M./Fehder, P.J. (eds)

> 1973 Data structures and accessing in data-base systems: Evolution of Information Systems. In: IBM Systems Journal, Volume 12, p. 30–44 [5].

Teehan, Aja/Gallagher, Damien/Keating, John

> 2008 The Digital Edition of the Alcalá Account Book Manuscript [1].

Vanhoutte, Edward

> 2006 Prose Fiction and Modern Manuscripts. Limitations and Possibilities of Text Encoding for Electronic Textual Editing. In: Lou Burnard, Katherine O'Brien O'Keeffe, John Unsworth (eds): Electronic Textual Editing. New York: Modern Language Association of America, p. 161–180 [6].

Websites

[1] <http://archives.forasfeasa.ie> (25.05.2009).

[2] <http://www.nist.gov/dads/> (25.05.2009).

[3] <http://cpsr.org/issues/pd/> (25.05.2009).

[4] <http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgann.xml#body.1_div.5> (25.05.2009).

[5] <http://www.minet.uni-jena.de/dbis/lehre/ws2005/dbs1/ibmsj1201C.pdf> (25.05.2009).

[6] <http://www.tei-c.org/About/Archive_new/ETE/Preview/vanhoutte.xml> (25.05.2009).

[7] <http://www.gimp.org/> (25.05.2009).

[8] <http://www.w3.org/XML/> (25.05.2009).

[9] <http://www.w3.org/XML/Schema.html> (25.05.2009).

[10] <http://www.uml.org/#UML2.0> (25.05.2009).

[11] <http://www.eprints.org/software/> (25.05.2009).

[12] <http://www.dspace.org> (25.05.2009).

[13] <http://www.fedora-commons.org> (25.05.2009).

[14] <http://www.encyclopedia.chicagohistory.org> (25.05.2009).

[15] <http://www.ucd.ie/ivrla/index.html> (25.05.2009).

[16] <http://www.openarchives.org/OAI/openarchivesprotocol.html> (25.05.2009).

[17] <http://dublincore.org/documents/dces/> (25.05.2009).

[18] <http://exist.sourceforge.net> (25.05.2009).

[19] <http://www.w3.org/TR/soap12-part1/> (25.05.2009).

[20] <http://www.artstor.org/what-is-artstor/w-html/features-and-tools-tech.shtml> (25.05.2009).

[21] <http://digitalgallery.nypl.org> (25.05.2009).

[22] <http://www.adobe.com/products/flex/> (25.05.2009).

[23] < http://maps.google.com/> (25.05.2009).

[24] <http://www.zoomify.com> (25.05.2009).

[25] <http://www.lib.umd.edu/digital/> (25.05.2009).

[26] <http://www.loc.gov/exhibits/> (25.05.2009).

[27] <http://www.gutenberg.org/etext/15114>(25.05.2009).

[1] Russell Library, Salamanca Archives, Legajo S30, nos 1–3.

[2] Vanhoutte provides this definition: »My full working definition of an electronic (scholarly) edition has six parts. By electronic edition, I mean an edition (1) which is the immediate result or some kind of spin-off product from textual scholarship; (2) which is intended for a specific audience and designed according to project-specific purposes; (3) which represents at least one version of the text or the work; (4) which has been processed in a platform-independent and non-proprietary basis, i.e. it can both be stored for archival purposes and also made available for further research (Open Source Policy); (5) whose creation is documented as part of the edition; and (6) whose editorial status is explicitly articulated.« (Vanhoutte 2006:161).

[3] »Participatory Design (PD) is an approach to the assessment, design, and development of technological and organizational systems that places a premium on the active involvement of workplace practitioners (usually potential or current users of the system) in design and decision-making processes«. [3]

[4] Users display varying preferences for means of accessing information; some prefer browsing over sequential access (as would be the norm for accessing a scroll) or category driven access (as in thematic access to a library's holdings through its catalogue). By facilitating both random and sequential access we are mimicking the access styles available to the user of the original manuscript. Any other access mechanisms, for instance keyword searches, are digital additions. For an overview of the evolution of random and sequential access in information systems please see Senko et al (1973).

[5] The Boolean operator in use is ›AND‹, which means that both of the search terms must be found within a page for that facsimile image to be displayed. For instance, if the keyword search is »bread wine« a page with just ›bread‹ will not be returned as a result. For further reading see [27].

[6] We used the GIMP software. More information can be found at: GIMP Developers: Homepage, [7].

[7] World Wide Web Consortium: Extensible Mark-up Language (XML), 1999, [8].

[8] World Wide Web Consortium: XML Schema, 2002 [9].

[9] The notion of »the work«, and all language-artefacts, as abstract, unchanging and thus unrealisable is a theme that underlies much of our two papers. For a discussion of the hermeneutics of a document please see J. McGann/D. Buzzetti: (2006).

[10] A hermeneutic spiral encapsulates the act of interpretation and understanding. Nothing that we examine is understood in isolation, but is rather understood in relation to what we know already and what it goes to compose. For instance, when we read a chapter in a book we have some notion of the book as a whole, which helps us to understand that chapter. This newly gained understanding then forms the basis for the next engagement – thus a spiral of understanding and interpretation is created. For an overview of the development of the concept please see Landa (2004).

[11] Trees are just one of a number of data structures such as networks, linked lists and stacks. For a description of data structures and some definitions, please see Black (2004).

[12] See section »Image Capture and Management« of this paper. A Use Case is a description of the interaction between a user and a system, along with the system's behaviour, during the response to a specific user request. For further reading see: Object Management Group: UML Specification, [10].

[13] Laurillard (2002).

[14] Fedora Object XML (FOXML) is a similar XML standard developed by Fedora for recording digital objects. Each digital object in a Fedora repository is stored internally as FOXML. Consequently, it is recommended by Fedora to use FOXML for ingestion of digital objects into Fedora repositories as the XML is a complete duplicate of how the digital object is stored in Fedora, whereas Fedora METS requires Fedora to perform internal translation in order to convert the digital object into FOXML.

[15] Metadata Encoding & Transmission Standard (METS) is a standard XML schema for describing the digital objects of a digital repository, their metadata and their administrative data within the repository. This common standard allows the ease of ingestion of digital objects amongst digital repositories. Fedora implements an extension of this standard known as Fedora METS, for recording Fedora specific features.

[16] Fielding/Gettys/Mogul/Frystyk/Masinter/Leach/Berners-Lee (1999).

[17] SOAP Version 1.2 Part 1: Messaging Framework (Second Edition), W3C Recommendation 27 April 2007, [19].

[18] Rapid Application Development is a software development methodology where models and prototypes are designed and built in rapid iterations, thus allowing for the incorporation of refined requirements. James Martin introduced the concept in 1991. For further reading see his book of that date.

[19] Dominick et al: (2000).