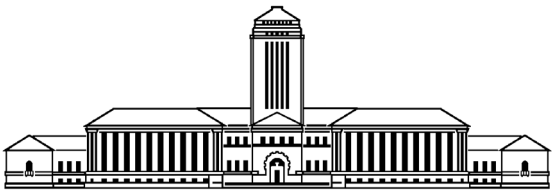

CAMBRIDGE UNIVERSITY LIBRARY

# Foundations Project

# Technical Plan (0.4)

**Summary**
This is a technical plan for the Foundations Project, setting out the functions we require from our new digital library infrastructure and the means by which we will achieve this.

## Contents



| Document Summary | |
|---|---|
| Title | Foundations Project: Technical Plan |
| Author | Grant Young |
| Version | 0.4 - Draft 4 |
| Date | 3-SEPTEMBER-10 |
| Distribution | Internal: Technical Group and Management Group |

# 1. Project Overview

## 1.1 Project background

Cambridge University Library (UL) wishes to make its superb collections available to serve the needs of those undertaking teaching & learning and research, within Cambridge and beyond. Digitisation is a key means of achieving this. Development of our digital infrastructure and digitised content is a key strategic objective of the UL.

Although the UL has undertaken significant digitisation, it can be characterised to date as ad-hoc and demand-driven rather than strategic. Some digitised content is available via the DSpace@Cambridge repository and the Library website, but much is currently in storage or is being delivered via the websites of other institutions or through commercial collections.

A scoping study was undertaken in 2007-09 to investigate support for image management for the UL and several of Cambridge's museums (CIP - Cambridge Images Project). This was funded by the Fitzwilliam and UL, and supported by CARET, Computing Services and Angela Murphy's consultancy. That project did much useful work, which we are likely to draw on in this project. However, whilst a CIP system would have been useful in aggregating and exposing images from across Cambridge's heritage institutions, the systems and approaches being proposed were unable to meet the full requirements of the UL, since they were concentrated on the management of individual images rather than the more complex digital objects the library needs to be creating and managing.

There were several attempts in 2008 and 2009 to raise money for a system to manage the library's digitised assets. In early 2010 we were successful in achieving £1.5m for a three year project to develop the necessary digital library infrastructure (c.£900) and undertake digitisation focused in the areas of Faith and Modern Science (c.£600k). In addition to this sum, the project will rely on some library contributions and has set itself a target of raising significant further funding towards digitisation and further development of the infrastructure.

## 1. 2. Project aims

The aims of the project as a whole are to:

- Provide a suitable **technical infrastructure** to support the creation, management and delivery of digitised content for Cambridge University Library;
- Create significant initial **digitised content** (individual items, collections and clusters) concentrating on the areas of Faith (Judaism, Christianity and Islam) and Early Modern Science (esp. Newton and the Astronomers Royal); and
- Provide a suitable **fundraising context** to enable us to attract further funding to complete the current project and extend it through further technical development and content creation projects.

## 1.3. Project scope

**Activities.** As indicated in the previous section, there are three broad areas of activity within the Foundations Project: (1) developing the technical infrastructure, (2) creating the content, and (3) fundraising for further work. Most of the project resources will be devoted to the first two activities and advisory groups will plan and inform each.

**Management.** The Foundations Project is led by a Management Group, which will report to the Librarian. The Management Group will oversee the work of other groups and will

manage the project's resources (budget and staff). It currently includes Patricia Killiard, Jill Whitelock, Ben Outhwaite and Grant Young.

**Timeframe.** The Foundations Project is expected to run from mid-2010 to mid-2013.

**Staffing.** The project will be able to fund posts to support its infrastructure development. In order to build on existing expertise and provide the project with greater sustainability, it will second some existing staff resources.

**Budget**. Approximately £900,000 can be spent on the development of the technical infrastructure over the three years of the project, including hardware, software, staff, consultancy and institutional overheads.

### 1.4. Project deliverables

The project intends to deliver the following main outputs:

- **Suitable technical infrastructure** to support the creation, management, preservation, delivery, reuse and enhancement of digitised content
- **Significant digital content** in the areas of Faith and Early Modern Science
- **Plans and resources to sustain and extend** the infrastructure and content beyond the initial three years of the project

## 2. Aims and values in developing the technical infrastructure

### 2.1. Aims

Our **broad aims** in developing the technical infrastructure are to:

- Provide a suitable technical infrastructure to support the ongoing creation, management and delivery of digitised content for the Library
- Ensure effective links with other infrastructure (metadata, discovery, learning and research environments, preservation)
- Ensure enough openness and interaction so that digital library content can be discovered, accessed and used via multiple channels and there is substantive feedback/contribution from users
- Enable us to efficiently process our legacy digitised content and metadata
- Enable us to further develop the infrastructure to manage an expanded range of content and functionality beyond the project
- Achieve good balances of quality, timeliness and cost efficiency

### 2.2. Values

As we develop the digital library infrastructure, we should be guided by the following values:

- **"Good enough" rather than "perfect"**
  We should aim to produce something of a high quality, but err on the side of pragmatism rather than perfectionism. It is very important that we deliver the project within time and budget so it may be that we have to deliver something "good enough" for now but provide enough flexibility to develop something "better" over time. "Good enough" in this context also means something that will satisfy the key stakeholders: funder, staff using the system, end users.

- **Most suitable tools for the job**
  There is no 'out of the box' solution – we will need to employ multiple technologies to achieve our aims. Although we budgeted for a central commercial system, we should remain open-minded about whether the solutions involve commercial products or are comprised largely or exclusively of open source software.
- **Solutions are not just technical**
  The infrastructure will rely on people – not all workflows can or should be automated. Many different people will play a role and must be actively involved and well supported.
- **Standards are important**
  Although standards in this area are still developing, where possible we should try to support best/good practices and existing and emerging standards. Any compromises should be deliberate rather than inadvertent and must be well justified and documented.
- **Interoperability and openness are important**
  The system must interact with multiple systems in terms of both its inputs and outputs. It needs to be as open as possible. Wherever possible, the content we create should be openly accessible and available for reuse within other systems and contexts
- **Flexibility and extensibility are important**
  We are unlikely to achieve everything we would wish within the scope of this project – and we should also expect the goal posts to shift over the duration of the project as technologies, standards and user expectations change. Our approach must be to develop something that can easily be adapted and extended.
- **Balancing the generic and the specific**
  We must try to achieve a good balance between generic and specific functionality. This is probably best done through providing tools and workflows that can be combined in different ways to the needs of particular projects, collections and content-types.
- **Scaleability**
  Our content will grow – we need to ensure that the infrastructure we put in place can accommodate that growth
- **Seperability**
  It will be useful to ensure that there is sufficient separation of content (metadata and associated digital objects) and infrastructure to enable them to exist and operate independently. The content will need to outlive its infrastructure and be easily moved as necessary.
- **Sustainability**
  While our goal will be to sustain and further develop the resource through other projects, this cannot be guaranteed. So whatever we produce will need to be sustainable beyond the project with fairly minimal effort.

# 3. Overview of requirements

## 3.1. High-level model

To state our requirements at the highest level, the technical infrastructure must support the **production** and **preservation** of and **access** to high quality digital **content** that provides a sufficiently good representation of the analogue originals to meet the needs and expectations of library users.

Figure 1 (below) models these main areas of activity and the primary flow of data. It is intended to provide a dynamic framework within which we can map the system and its

various workflows, technologies and interactions. It places digital content at the heart of the system, but recognises that production, preservation and access activities are all going to be critical to its success.

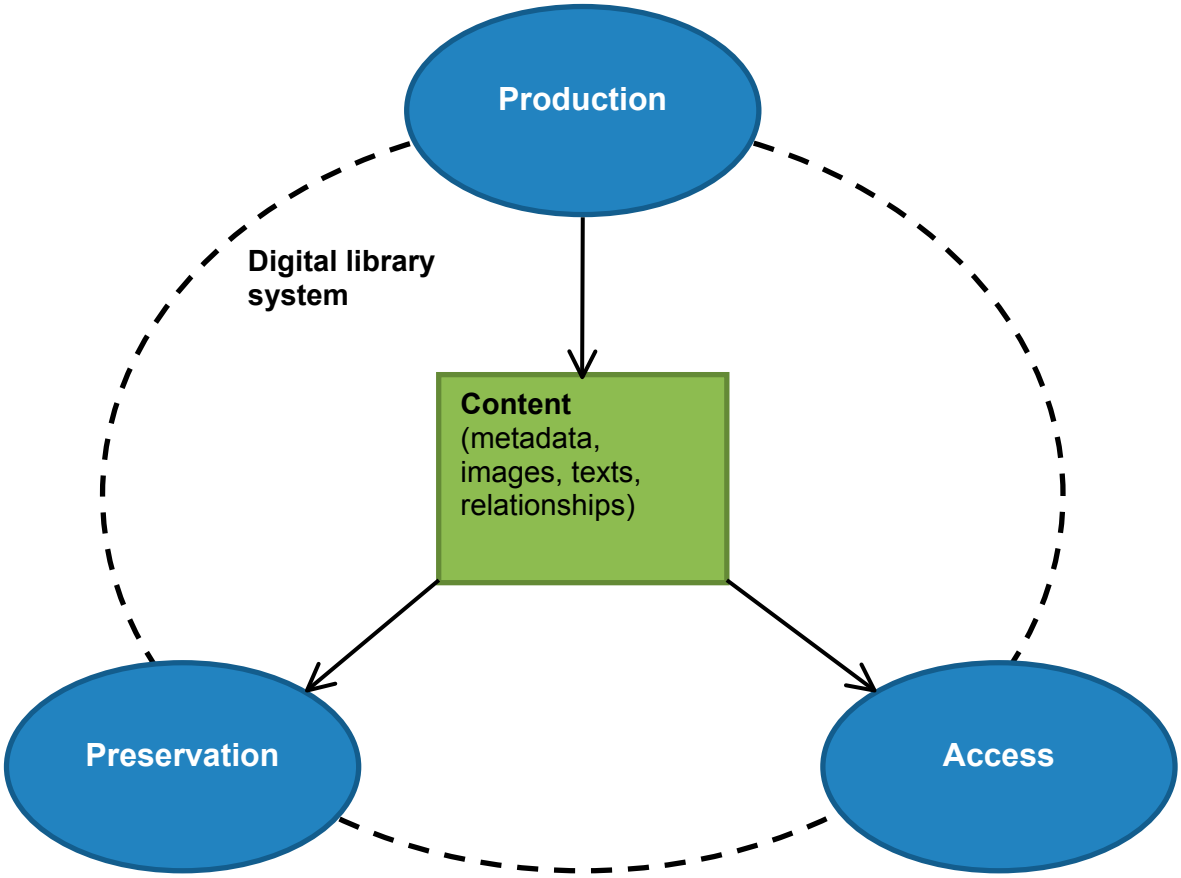


*Figure 1. High-level overview of the digital library system*

Note that the activities overlap the boundaries of the digital library system – indicating that production, preservation and access are not the exclusive preserve of this infrastructure. **All these activities interface with other systems**. For example:

- in terms of **production**, the system will need to interface with the library management system (Voyager) and archives management system (Cantab/Janus) to acquire descriptive metadata, and with digitisation software to acquire images and technical metadata;
- in terms of **preservation**, it will need to interface with storage systems and the university's digital repository (Dspace@Cambridge) to manage and preserve the data and metadata;
- in terms of **access**, it will need to interface with the resource discovery system (Aquabrowser) to enhance discovery, and with custom interfaces managed by the library or external parties (via APIs) to enhance the delivery and use of the content.

Note too that while the arrows indicate the main flows of data, there are **further flows and cycles** within this system. It is anticipated, for example, that:

- the **access** activity will not just provide a passive presentation of content, but will itself generate further content, through user contributions – so it feeds back into the **production** activity.
- Similarly, **preservation** should not be envisaged as the one-time passive storing of bit-streams, but an actively managed process that will result in format migrations and

the updating of metadata – i.e. it will feed back into the **production** of further/revised content.

Figure 2 (below) illustrates these interactions and secondary flows of data.

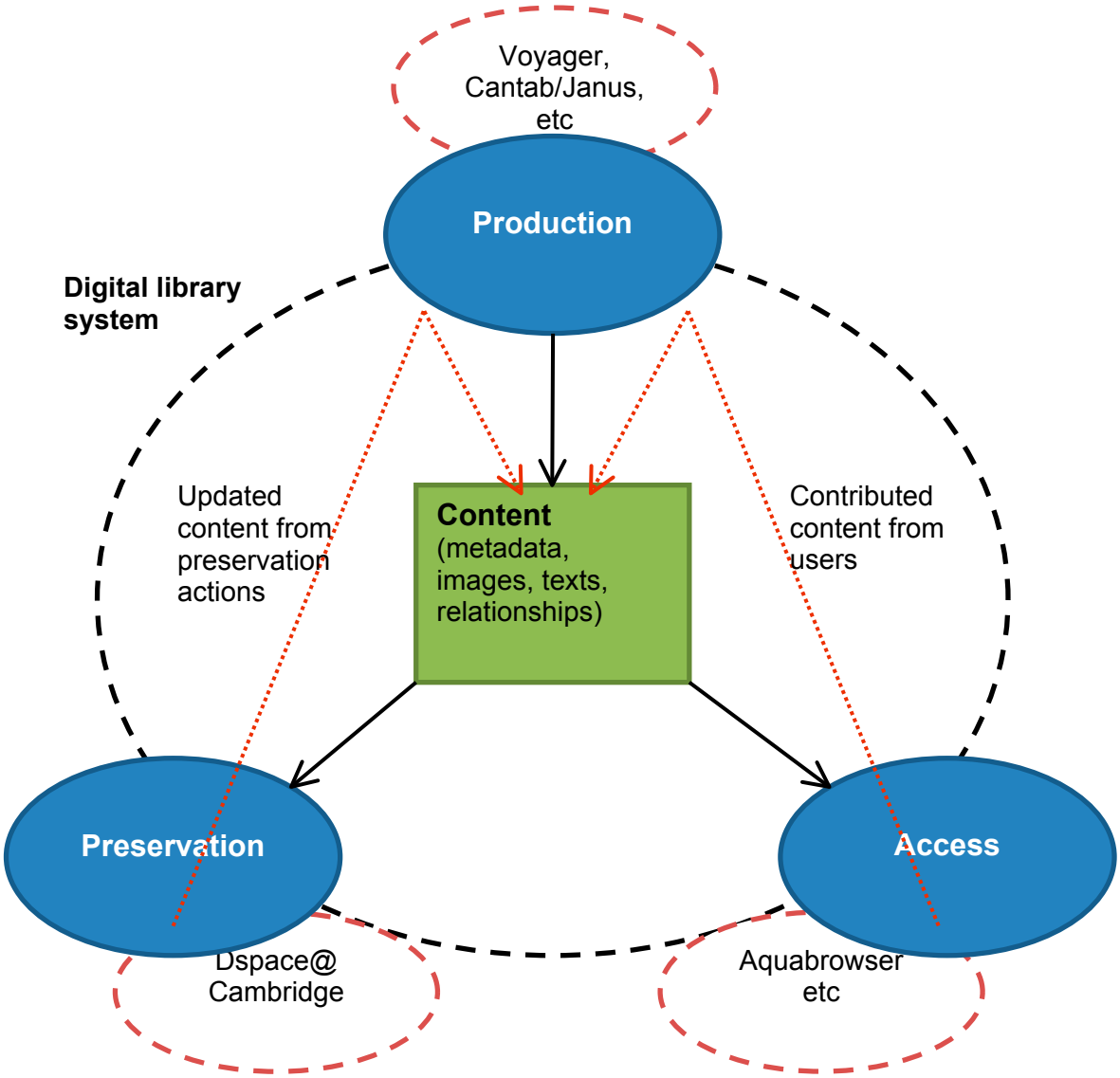


*Figure 2. Examples of interaction and feedback*

## 3.2. High-level description of requirements

***Overview***

- The system should be able to meet the needs of a wide range of **users** – both content producers and consumers, including digitisation staff, collection curators, experts, researchers, teachers, learners and the interested public.
- It should to be able to manage and deliver a wide variety of digitised **content**, particularly texts, artistic and photographic works, but also objects, moving images and sound.

- It should provide efficient **interoperability** with several other systems, including the library and archive management systems, resource discovery system, DSpace@Cambridge repository, and CamTools, and it should open the content to further discovery and reuse.
- It should support several **workflows**, including digitisation, metadata creation and transcription, copyright, licensing and sales, preservation, user enhancement and reuse of content…
- It should offer a good generic **interface** to the content along with the ability to produce more tailored presentations to suit particular items, collections or projects, including the creation of formal digital publications/editions.
- It should offer sophisticated **resource discovery**, making full use of available metadata and text and providing multiple pathways through the content.
- It should offer a high level of user **interaction**, enabling the user to manage their use of the digital library and its contents (e.g. accessibility, bookmarking), to contribute content (e.g. annotations, comments, metadata, transcriptions, relevant links), and where permissible reuse the metadata or content in other contexts.
- It should provide sufficient **sustainability**, so that it can be maintained with minimal resources beyond the project, but also the **flexibility** to be developed further should resources or new technologies permit this.

***Users***

- The system should provide good support for internal/administrative users (e.g. digitisation project staff, curators) and external/'end' users (e.g. researchers, teachers, students, whether at Cambridge or beyond).
- The internal/external user distinction is not necessarily clear-cut and does not exactly correspond to producer/consumer. We want to provide an environment and tools to enable users to further enhance our content – in formal ways (e.g. an online critical edition using TEI) and less formally (e.g. a wiki discussion around an important manuscript).
- It is also likely that some of the internal/administrative users will be based beyond the UL, Cambridge, and the UK.
- Some typical use cases are described in section 4 below.

***Content***

- The system must be able to present the full range of library content, which is very diverse. A large proportion of the content will take the form of complex digital objects: scans of pages, suitably sequenced, with associated metadata at various levels (whole work, parts, pages, parts of pages) and in some cases associated texts (OCR or transcription) and other information (contextual information, bibliographical references, links, user contributed content…). There will also be single images: photographs, artistic works, individual leaves and fragments. Some of the content will form collections or sub-collections (e.g. Genizah fragments); others will stand alone as individual items (e.g. the Gutenberg Bible, the Nash papyrus).
- Although the focus for content creation in this phase of the Foundations Project is on manuscripts related to three faiths of the book (Judaism, Christianity and Islam) and early Modern Science (Newton and the Astronomers Royal), we will also need to accommodate existing digitised content (which we will be evaluating early in the project) and the on-going demand-driven digitisation undertaken by Imaging Services. It is likely that we will also pick up additional funded projects over the next three years that do not fit into the Faith/Science framework but will need to make use of this platform.
- Our intention is that master versions of our digitised resources will reside in the DSpace@Cambridge repository, with suitable delivery surrogates in the digital library system.

- The system should support a variety of ingest mechanisms – individual upload and batch upload, possibly a content 'drop box' model or the ability to pull and transform content on the fly from another system. Depending on how this digital library system is related to DSpace@Cambridge, the typical ingest route might either be direct from Imaging Services or via the repository.
- In addition to digitised content, the system should be able to handle a variety of metadata: bibliographic, transcription mark-up, technical, preservation-related, rights-related, user-contributed data. It should be able to handle XML imports and exports, mapping as appropriate.

***Interoperability***

- This system cannot sit in isolation and must work as effectively as possible within a broader set of systems.
- Much metadata exists in the library management system (Voyager/Newton) and the archival management system (Cantab/Janus). Metadata can also be generated by software used in the digitisation workflow. The system must exploit this data.
- DSpace@Cambridge is our preferred preservation repository, where the master/archival digital content will reside. We will need to ensure that data is passed between these systems. It may be that the content goes first into DSpace@Camrbidge and is then pulled into the digital library system for metadata addition and delivery. Or it may be that the system manages the files until they are ready to be published and then writes the large file and archival metadata to DSpace. This workflow will need some further thought and discussion.
- The UL is implementing a resource discovery system to provide search and discovery across a wider range of library databases. Content from the digital library system must be able to be discovered via this means. The system should also enable discovery via external services, such as Google. Its metadata records should be harvestable and should also be available as feeds.
- We also want to see the content widely used and reused, so the ability to discover and pull/push content within Camtools, library or departmental websites, external aggregations etc is desirable. This may be at an item level, or at a collection level – e.g. if we wanted to contribute standardised data in batches to external content aggregators.

***Workflows***

- ***Digitisation.*** The system should offer good support for the Imaging Services unit, enabling them to fairly effortlessly upload content they are creating and attach relevant technical or descriptive metadata. Ideally it should also support quality assurance activities, reporting, and some of the commercial activities of this unit. There will need to be some sort of ordering/request process to enable a user to put in a request for a high-resolution version of an image in the system for publication. Online purchasing would be nice to have but is probably out of scope for this project.
- ***Metadata and transcription.*** While metadata exists for many of the items in our collection, it will typically require transformation and the addition of further metadata before the resources are published online. Some of this additional metadata may be automatically generated (e.g. technical or structural metadata), but some may need to be added by curators and experts (who may be internal or external to the UL and Cambridge). There is interest in transcribing many of our collections and external transcriptions already exist for much of the Newton material (in TEI) and Darwin material (bespoke mark-up) and some other items (e.g. Scriptorium project). It may be that some of these existing transcriptions can be brought into the system in the future, but we need to be providing tools to support scholars in creating such transcriptions within the system. It will be important that the system can handle non-English and bi-directional character sets.

- ***Rights/licensing.*** While our ambition is to provide openly accessible content, we will not always be able to offer the content under open licences. Many of our recent manuscript collections are still in copyright, which is owned by third parties. We may receive permission to mount content openly, but are unlikely to receive permission to license it via, for example, Creative Commons. The system needs to enable us to record information about rights and licences – for our own internal management and in order to provide appropriate notices and access-/use- permissions to end users.
- ***Sales.*** As noted above, we need to be able to support Imaging Services in ordering of high resolution versions for publication. This is most likely to be based on forms.
- ***Preservation.*** The DSpace@Cambridge repository is the university's preservation service and we expect its preservation functionality to be enhanced in future years. The digital library system must ensure that preservation versions of its content and metadata are deposited into DSpace@Cambridge. Given our focus on user interaction and contribution, it may be necessary to make some differentiation of content: with some content scheduled for long-term preservation; and other content regarded as 'potentially disposable.'
- ***User interaction and enhancement.*** The system should provide good support for users in their use of the resource. This might include customisation, the ability to sign up to content feeds, annotate content, add information within a wiki environment, provide additional metadata, or add links to related resources and scholarship.

***Interface***

- The system needs to offer a generic interface capable of presenting any type of content and enabling a full search across all content. But it also needs to provide and enable more tailored presentations, according to content type, collection and project. We will need to be able to "brand" particular content according to funder, and create themed collections, exhibition sites, and online editions. The interface needs to be visually attractive, accessible to those who have impairments or are accessing content on alternative platforms, and be highly functional/usable.
- It will probably be sensible to maintain a separation between the interface and management layers. If we purchase a commercial system, it may be that we can rely on this for a generic delivery and develop more specialised interfaces on top of it.
- We would like to enable interaction with the resource in addition to passive consumption. Depending on the content, this might range from enhancement by approved scholars to crowd-sourcing, user tagging, and wiki features. Some distinction may need to be made between content that is 'authorised' and 'core' and other content that is 'contributed'.

***Resource discovery***

- As noted above, we need to provide good resource discovery within the system and via other services operated by the library or externally. A goal will be to enable individual resources to appear within a general Google search.
- The system should provide good search facilities, along with the ability to narrow down the searching by facet. There should also be pathways across the content (e.g. hyperlinked metadata) so the user does not reach dead-ends or need to return to a main search or high-level browse to reach further content.
- The library has a lot of material in non-English languages and scripts, so it will be important to support these in searching and displaying content. The initial Foundations of Faith collection is likely to include Hebrew, Arabic and Greek text, and further digitisation is likely to include East Asian scripts. These will all pose challenges.
- It will be important to provide addressable persistent identifiers for resources, so they can be reliably bookmarked and linked to. It may be advantageous to pass links back

to the library catalogue, although the new resource discovery platform will offer some assistance in relating physical items to their digitised surrogates.

***Interaction***

- Previous sections have indicated that the system should support a high degree of interaction. This may include in-system functionality or Web 2.0 “add-ons”. It may be that some of the interaction needs to be phased in its delivery or that it is offered via bespoke interfaces rather than provided as core functionality.

***Sustainability and flexibility***

- We have secured three years of funding, but cannot yet be certain of resourcing beyond that period (i.e. from mid-2013). Our goal will be to sustain and further develop the resource through further development projects or digitisation projects, but this cannot be guaranteed. So whatever we procure or build will need to be sustainable beyond the project with minimal effort.
- Nor can we predict with certainty the emergence of new standards and technologies or what will happen to our adjacent infrastructure (library system, archives system, repository, resource discovery system..). So the system we put together should also offer sufficient flexibility to “add-on/in” new functionality, swap out components, or interoperate with alternative systems with minimal effort.

## 4. Typical use cases

*To be developed*

## 5. Detailed specification

This section lists specific support and functionality we require or desire within a digital library system. **It is important to note that not all of these requirements can be met within a single system or development, nor will all be possible within a three-year project.** The priority column gives an indication of the importance of the requirement (high, medium, low) or indicates that information about this must be sought about any core system (Info required). In the notes field we have tried to indicate where a feature might be expected within the core system or addressed through additional software or development.

This list has benefited from recent specifications produced by the SAFIR project at the University of York (2008) and the ITT for digital object repository by the Wellcome Trust (2008), although each was specifying a different kind of system from ours (York, a general repository infrastructure; Wellcome, a preservation repository).

| Ref | Requirement | Description | Priority | Notes |
|---|---|---|---|---|
| **1** | **Production of content** | | | |
| **1.1** | **Managing the production** | | | |
| 1.1.1 | Production scheduling and monitoring | Ability to schedule and monitor production of digital content at item level or project level | High | We expect this within core system |
| 1.1.2 | Production staff management | Ability to assign tasks to imaging staff and monitor progress | High | We expect this within core system |
| 1.1.3 | Quality Assurance (QA) for images | Tools/metadata for scheduling, undertaking and recording quality checks of images | High | We expect this within core system |
| 1.1.4 | Quality Assurance (QA) for OCR text | Tools/metadata for scheduling, undertaking and recording quality checks of OCR text | Medium | Nice to have - we will need to investigate the availability and maturity of these tools |
| 1.1.5 | Quality Assurance (QA) process for metadata | Tools/metadata for scheduling, checking and recording quality of metadata assigned to items | High | May need to be addressed through another tool or development work |

| | | | | |
|---|---|---|---|---|
| 1.1.6 | Quality Assurance (QA) for transcriptions, annotations, commentary | Tools/metadata for scheduling, undertaking and recording quality checks of transcriptions, annotations, commentary associated with items | High | May need to be addressed through another tool or development work |
| 1.1.7 | Production information | Tools/metadata for recording information about the production (producer, capture equipment, date etc) | High | We expect this within core system |
| 1.1.8 | Production reporting | System must generate a range of production reports to support production management | High | We expect this within core system |
| **1.2** | **Ingesting content** | | | |
| 1.2.1 | Ingest and processing of existing content | Ability to ingest and process pre-existing sets of images, texts, metadata in batch processes | High | We expect this within core system |
| 1.2.2 | Integration with capture hardware and software | Ability to obtain content directly from capture or processing software | Medium | It may be useful to have this within the core system |
| 1.2.3 | Image formats | Can ingest and handle a wide range of image formats including RAW, TIFF, JPEG, JPEG2000, GIF, PNG, BMP, PDF | High | We expect this within core system |
| 1.2.4 | Other media formats | Ability to ingest and manage sound and moving image files | Medium | It may be useful to have this within the core system |
| 1.2.5 | Metadata formats | Can import/map-in descriptive metadata in variety of forms including MARC21, MARCXML, Dublin Core, MODS, EAD, TEI header, bespoke mark-up | High | We expect this within core system |
| 1.2.6 | Text formats | Can import text in a variety of forms (OCR data, TEI mark-up, bespoke mark-up) | High | We expect this within core system |

| | | | | |
|---|---|---|---|---|
| **1.3** | **Image production** | | | |
| 1.3.1 | Image manipulation and processing | Provides control over pixel dimensions, filesize, bit-depth, colourspace and compression, ability to perform rotation, crop, brightness/contrast and hue/saturation adjustments. Individually or in batches | Medium | Might be provided by imaging tools instead of core system |
| 1.3.2 | Colour management | Ability to embed or associate profiles and use them in image processing | Medium | Might be provided by imaging tools instead of core system |
| 1.3.3 | Watermarking | Ability to apply visible or invisible watermarks | Medium | Nice to have if available within core system. Alternatively might be delivered on the fly to end user |
| 1.3.4 | Marking regions of interest on images | Tools for marking regions on images (e.g. using JPEG2000) | Low | Nice to have if available within core system or another tool |
| 1.3.5 | Surrogate generation | Creation of digital surrogates according to custom specification, individually or as batched process | High | We expect this within core system, but might alternatively be generated on the fly for export or end-user access |
| 1.3.6. | Metadata extraction | Extraction of technical and other metadata from file headers | High | Might be provided by imaging tools instead of core system |
| 1.3.7 | Metadata embedding | Embedding of metadata (e.g. descriptive or rights) in digital files | Medium | Might be provided by imaging tools instead of core system |
| **1.4** | **Text production** | | | |
| 1.4.1 | OCR of text-based images | OCR tools, enabling plain text and recording of coordinates for highlighting | Medium | We expect this within core system or associated tools. OCR is not high priority for Foundations Phase 1 material, but will be important for other content |
| 1.4.2 | Layout mark-up of OCR text (ALTO standard) | Tools for recording mark-up (e.g. ALTO standard) | Low | Nice to have, although ALTO is not yet widely supported |

| | | | | |
|---|---|---|---|---|
| 1.4.3 | Transcription | Templates to support production of transcriptions (TEI and bespoke) | High | Likely to be provided by other tools or development work rather than core system |
| **1.5** | **Metadata production** | | | |
| 1.5.1 | Image filenaming support | Tools for naming and renaming files (individually and in batches) | High | We expect this within core system |
| 1.5.2 | Descriptive and contextual information | Tools/metadata for adding or enhancing description of item or collection | High | We expect this within core system |
| 1.5.3 | Controlled vocabulary support | Ability to manage and assign controlled terms, including tags, subject keywords and thesaurus terms | High | We expect good support within core system but may need additional tools or development |
| 1.5.4 | Georeferencing | Tool to assign georeferences to items | Medium | Nice to have - we will need to investigate the availability and maturity of standards and tools |
| 1.5.5 | Conservation and handling information | Tools/metadata to tag and annotate item records with information about condition and handling of original material | High | Likely to be a customisation or development of core system |
| 1.5.6 | Copyright and licensing information | Tools/metadata for recording rights information about (copyright status, clearance information, licensing conditions) | High | Likely to be a customisation or development of core system |
| 1.5.7 | Sales and ordering information | Tools/metadata for recording sales-related information | High | Likely to be a customisation or development of core system |
| 1.5.8 | Bibliographic data | Tools for creating and managing references related to items in the collections (citations, websites, other resources etc) | High | Nice to have if available within core system, but likely to require management through other tools |

| | | | | |
|---|---|---|---|---|
| 1.5.9 | Spell-checking | Ability to check metadata and texts against dictionaries | Medium | We expect this within core system |
| 1.5.10 | Image filenaming support | Tools for naming and renaming files | High | We expect this within core system |
| 1.5.11 | Validation of metadata | Ability to constrain data entry and check XML encoding against schemas and DTD | High | We expect this within core system |
| **2** | **Management of content** | | | |
| **2.1** | **Identification and organisation of content** | | | |
| 2.1.1 | Unique and persistent identifiers | Ability to assign unique, persistent identifiers to digital files | High | We expect this within core system |
| 2.1.2 | Support for identifier schemes (e.g. handles or DOIs) | Ability to assign and use identifiers according to standard schemes | High | We would like this within core system |
| 2.1.3 | Multiple related manifestations/versions | Ability to store and relate multiple manifestations and versions of same digital object (e.g. different filesizes, colourspaces) | High | We expect this within core system, but might alternatively be generated on the fly for export or end-user access |
| 2.1.4 | Multiple related formats | Ability to store and relate mulitple formats for same digital object | High | We expect this within core system, but might alternatively be generated on the fly for export or end-user access |
| 2.1.5 | Grouping and sequencing of digital objects | Ability to relate digital objects to form sequenced items | High | We expect this within core system |
| 2.1.6 | Support for collections and hierarchies | Ability to group items into collections and express hierarchical relationships (in ways other than folder structures) | High | We expect this within core system |
| **2.2** | **Hardware, infrastructure and operating system requirements** | | | |

| | | | | |
|---|---|---|---|---|
| 2.2.1 | Client requirements | Please describe hardware and operating system requirements for any clients, web or desktop based | Info required | Web client strongly preferred |
| 2.2.2 | Server requirements OS | Please describe supported and prefered server operating systems | Info required | |
| 2.2.3 | Hardware requirements Server | Given our expected user base, please detail the recommended server specification (with suggested make and specification of hardware for a customer with a user base similar to Cambridge). Please also include details of specific equipment used by existing customers | Info required | |
| 2.2.4 | Efficient storage of content | We require storage in robust scalable environment separate from system components. Describe how this can be achieved with examples from existing users | Info required | |
| 2.2.5 | Large file sizes | We require the ability to manage large files (up to 500MB) and volumes. Are there known limits? | High | We expect this within core system |
| 2.2.6 | Virtualisation | What virtualisation environments do you support? Are any customers virtualising already? What is required? | Info required | |
| 2.2.7 | Clustering and load balancing | How does the solution support clustering and load balancing of computational tasks and web server management | Info required | |

| | | | | |
|---|---|---|---|---|
| 2.2.8 | Network access | We would like the ability to access and manage files on/from any network location | Info required | |
| **2.3** | **Software requirements** | | | |
| 2.3.1 | Application architecture | Is system client or Web-based? | Info required | Web client strongly preferred |
| 2.3.2 | Programming languages used | What programming languages and environments are used by the product? | Info required | |
| 2.3.3 | Supported databases | What databases does system require/support? Use of open industry standards preferred | Info required | We need to discover whether metadata is primarily stored as XML or within database |
| 2.3.4 | Indexing system | What indexing mechanisms are used by the product. How often are they updated / refreshed? | Info required | |
| 2.3.5 | Character support | Support for UNICODE; ability to handle non-Latin character sets and alternative orientations (e.g. right-left) | High | We expect this within core system |
| 2.3.6 | Other software issues | Is there any other advantage or problem or disadvantage in the System, of which the University should be aware? | Info required | |
| **2.4** | **Systems management and performance issues** | | | |
| 2.4.1 | Upgrades | What is your policy with regard to developments, including upgrades, enhancements, major developments? How frequently are these provided? | Info required | |
| 2.4.2 | Availability during upgrades | Please describe how system upgrades are handled. What level of downtime is usually involved? | Info required | |

| | | | | |
|---|---|---|---|---|
| 2.4.3 | Enhancement requests | Please describe how enhancement requests from customers are handled, assessed and prioritised into your development schedules | Info required | |
| 2.4.4 | Releases of new versions | Please describe how frequently new versions are released. What constitutes a major release. In addition, how often are regular point releases or patches produced. How optional are these | Info required | |
| 2.4.5 | Backup and recovery | Please describe any facilities and/or requirements for backup and recovery. What options are existing customers making use of? | Info required | |
| 2.4.6 | Performance and benchmarking | Please provide performance details and benchmark figures for any maximum concurrent user testing | Info required | |
| 2.4.7 | Capacity limitations | Please provide details on the theoretical maximum amount of records and objects storable, and on the maximum tested and currently operational records stored by the system | Info required | |
| 2.4.8 | System monitoring | Please describe options available for monitoring of systems processes and tasks. | Info required | |
| **2.5** | **Access control and security** | | | |

| | | | | |
|---|---|---|---|---|
| 2.5.1 | Single sign-on integration | Please provide information on any potential for interaction with university single sign-on services (staff side administration and end-user public facing interfaces) | Info required | |
| 2.5.2 | Federated access control | Please provide information on any interaction with the Shibboleth/SAML based federated access management services | Info required | |
| 2.5.3 | Data security (SSL) | | Info required | |
| 2.5.4 | Password security | Please detasil how passwords are stored and entered into the system | Info required | |
| 2.5.5 | Antivirus support/integration | Please detail options for integrating antivirus mechanisms with file storage | Info required | |
| 2.5.6 | Integrity checking/validation | Systems to check the integrity of the data held in the system (e.g. checksums) | Medium | We expect this within core system |
| 2.5.7 | Auditing | Must provide good history of activity | High | We expect this within core system |
| **2.6** | **System integration** | | | |
| 2.6.1 | Integration with email | Ability to link system to email to enable notifications for staff | Info required | |
| 2.6.2 | Integration with Request Tracker | Ability to link system to Request Tracker to provide support | Info required | This is likely to require local development/integration |
| 2.6.3 | Integration with Voyager | Ability to integrate with Voyager Library management system to source bibliograhic and holdings information at various stages, both for individual works and e-mass as a batch process | Info required | This is likely to require local development/integration |

| | | | | |
|---|---|---|---|---|
| 2.6.4 | Integration with Cantab/Janus | Ability to integrate with Cantab/ Janus archives management system to source bibliograhic and holdings information at various stages, both for individual works and e-mass as a batch process | Info required | This is likely to require local development/integration |
| 2.6.5 | Integration with Camtools/Sakai | Export and import of objects to Sakai environment - embed as learning objects | Info required | This is likely to require local development/integration |
| 2.6.6 | Integration with DSpace | Please see section 3 for more details | Info required | This is likely to require local development/integration |
| **2.7** | **Licensing and code deposit** | | | |
| 2.7.1 | Source code licensing | What license(s) are the source code published under. Who owns any rights to the use of source code? | Info required | |
| 2.7.2 | Escrow | Has the source code been deposited under an ESCROW agreement? | Info required | |
| **3** | **Preservation of content** | | | |
| **3.1** | **Deposit/Ingest** | | | |
| 3.1.1 | Transfer of master images to preservation repository | Transfer of master image files to DSpace@Cambridge for long-term storage | High | This is likely to require local development, but will be dependent on core system APIs/exports |
| 3.1.2 | Transfer of archival metadata to preservation repository | Transfer of metadata to DSpace@Cambridge as XML files (METS, TEI and associated marked up files) | High | This is likely to require local development, but will be dependent on core system APIs/exports |
| 3.1.3 | Transfer of archival text to preservation repository | Transfer of OCR or transcribed text | High | This is likely to require local development, but will be dependent on core system APIs/exports |

| | | | | |
|---|---|---|---|---|
| 3.1.4 | Generation of repository metadata | Mapping from METS to DSpace@Cambridge Dublin Core profile for management and repository discovery | High | This is a local development |
| **3.2** | **Access** | | | |
| 3.2.1 | Access to low-resolution repository surrogates | Production (on ingest or on fly?) of low-resolution versions for access | Medium | This is a DSpace development |
| 3.2.2 | Access to high-resolution masters | Access to masters by approved staff/systems for order fulfilment, high-end uses (e.g. printing) or re-ingest into digital library system | High | This is a DSpace development |
| **3.3** | **Preservation management** | | | |
| 3.3.1 | Preservation tools | To be further developed within the context of DSpace@Cambridge. Must support active preservation and conform to the requirements of OAIS. | Medium | This is a DSpace development |
| **4** | **Access to content** | | | |
| **4.1** | **Interface** | | | |
| 4.1.1 | Generic digital library interface | General interface enabling browse, search and rendering of all content | High | We expect this within core system |
| 4.1.2 | Tailored interfaces | Custom interfaces to particular items, collections, content types | High | We expect core system to enable multiple customised interfaces |
| 4.1.3 | Accessibility of interface | Interfaces must meet accessibility standards/guidelines (e.g. W3C) | High | We expect this within core system |
| 4.1.4 | Support for standard browsers | Good access across a range of common browsers and versions | High | We expect this within core system |
| 4.1.5 | Support for multiple devices | Good access on multiple devices including mobile devices | Medium | We expect this within core system |
| **4.2** | **Search and discovery** | | | |
| 4.2.1 | Single configurable search prompt | Google-like single search on defined fields/elements | High | We expect this within core system |

| | | | | |
|---|---|---|---|---|
| 4.2.2 | Advanced configurable search | Advanced searching, including fielded searching, stemming and Boolean operators | High | We expect this within core system |
| 4.2.3 | Synonym and spelling suggestions | Prompts where search results are zero or low | Medium | Nice to have if available within core system |
| 4.2.4 | Browsing of content based on wide range of facets | Browsing from start-page, search results and at item level | High | We expect this within core system |
| 4.2.5 | Amazon-like prompts | Prompts based on previous activity of user or similar users (e.g. "others who viewed this looked at...") | Low | Nice to have if available within core system |
| 4.2.6 | User access to controlled vocabularies | Vocabularies available to end users to support searching and browsing | Medium | Nice to have if available within core system |
| 4.2.7 | User tagging | User-contributed tags can be used in browsing | Medium | Nice to have if available within core system |
| 4.2.8 | Date range and timeline searching and browsing | Use of dates to provide date-range searches and timeline presentations | Medium | Nice to have if available within core system |
| 4.2.9 | Map-based searching and browsing | Use of georeferences to provide map-based presentations | Medium | Nice to have if available within core system |
| 4.2.10 | Support for UNICODE | Search and display in UNICODE, including right-to-left orientation | High | We expect this within core system |
| 4.2.11 | Indexing by Google | Content is indexed by Google at page level and highly placed | High | Nice to have if available within core system. Otherwise we will have to achieve via other means |
| **4.3** | **Structuring of content** | | | |
| 4.3.1 | Support for collections | Ability to group and view content within multiple collections | High | We expect this within core system |
| 4.3.2 | Support for hierarchies | Ability to represent archival hierarchies | High | Nice to have if available within core system. Otherwise we will have to rely on other systems (e.g. Cantab/Janus) or undertake development |

| | | | | |
|---|---|---|---|---|
| 4.3.3 | Support for complex objects | Ability to display multi-page and multi-sided objects in order | High | We expect this within core system |
| 4.3.4 | Transcriptions alongside page images | Ability to display transcriptionsalongside images, preferably with multiple rendering options (e.g. diplomatic, semi-diplomatic) | High | Nice to have if available within core system. Otherwise we will have to achieve within custom interfaces |
| 4.3.5 | Support for 'online critical editions' | Ability to create formal editions with facsimiles, transcriptions, translations, critical appartus | Medium | Nice to have if available within core system |
| 4.3.6 | Support for 'online exhibitions' | Ability to create exhibitions with captions, commentary and learning resources | High | Nice to have if available within core system. Alternatively may be supported through other tools (e.g. OMEKA) or local development |
| **4.4** | **Image display and manipulation** | | | |
| 4.4.1 | Image viewer - panning | Ability to move across the image when displayed at high resolution | High | We expect this within core system |
| 4.4.2 | Image viewer - zooming up to 100% | Ability to expand up to (but not beyond) 100 percent of the largest image | High | We expect this within core system |
| 4.4.3 | Image viewer - rotation | Ability to turn the image | Medium | Nice to have if available within core system |
| 4.4.4 | Image viewer - comparison between images | Ability to compare one or more images | Low | Nice to have if available within core system |
| 4.4.5 | Image viewer - manipulation | Ability to make adjustments to the image (e.g. to its colour or contrast) | Low | Nice to have if available within core system |
| 4.4.6 | Streaming of JPEG 2000 images | Using JPEG2000 to provide zoom | Medium | Nice to have if available within core system |
| 4.4.7 | Page turner - swapping images | Images change when next page selected | High | We expect this within core system |

| | | | | |
|---|---|---|---|---|
| 4.4.8 | Page turner - graphic flipping | Images seem to flip when next page selected (e.g. Turning the Pages) | Medium | Nice to have if available within core system |
| 4.4.9 | Page turner - jump to specific page | Ability to select a page to view within a large volume | High | We expect this within core system |
| 4.4.10 | Surrogate generation on the fly | Automatic generation of images | Medium | Dependant on system |
| **4.5** | **Downloads, feeds, exports** | | | |
| 4.5.1 | PDF downloads | Ability for end user to download a PDF of the entire object | Medium | Nice to have if available within core system |
| 4.5.2 | Page image downloads | Ability for end user to download individual pages | Medium | Nice to have if available within core system |
| 4.5.3 | Metadata downloads | Ability for end user to obtain bibliographic records | Medium | Nice to have if available within core system |
| 4.5.4 | Generation of RSS feed | RSS to enable easy syndication | Medium | Nice to have if available within core system |
| 4.5.5 | Generation of RDF | RDF encoding to support semantic web applications | Medium | Nice to have if available within core system |
| 4.5.6 | Expose as OAI data | Expose records in OAI compliant form for harvesting and resource discovery | High | We expect this within core system. Alternatively we could rely on AcquaBrowser |
| 4.5.7 | Export as RDF-encoded metadata record | To support semantic web applications | Medium | Nice to have if available within core system |
| 4.5.8 | Export as TEI record | | High | Nice to have if available within core system, but may require development or another tool |
| 4.5.9 | Export as Dublin Core | To support discovery and aggregation and transfer to Dspace@Cambridge | High | We expect this within core system |
| 4.5.10 | Export as MARC record | For potential incorporation into catalogue | High | We expect this within core system |
| 4.5.11 | Export as MODS record | | High | We expect this within core system |

| | | | | |
|---|---|---|---|---|
| 4.5.12 | Export as METS with wrapped metadata (MODS, DC, MIX and PREMIS) | To support transfer to Dspace@Cambridge | High | We expect this within core system |
| 4.5.13 | Export as METS with embedded OCR | | Medium | Nice to have if available within core system |
| 4.5.14 | APIs | APIs to enable library and external parties to develop interfaces to the digital library content | High | We expect this within core system |
| **4.6** | **Access and rights** | | | |
| 4.6.1 | Access management - define by network location | Manage access at domain level | Medium | We expect this within core system, but may have to undertake development or use other tools |
| 4.6.2 | Access management - define by groups/roles | Manage access by group membership | Medium | We expect this within core system, but may have to undertake development or use other tools |
| 4.6.3 | Access management - define individuals | Manage access at individual level | Medium | We expect this within core system, but may have to undertake development or use other tools |
| 4.6.4 | Visible watermarking of content | Watermarks applied as part of production process or on the fly to selected content | Medium | Nice to have if available within core system |
| 4.6.5 | Personalised watermarking of content | System stamps image with date/time and user information | Low | Nice to have if available within core system |
| 4.6.6 | End-user rights information | Rights statements/licenses at collection and item levels, display of CC licences where appropriate | High | We expect this within core system, but may have to undertake development |

| | | | | |
|---|---|---|---|---|
| 4.6.7 | End-user rights management | Limitation of deliverables, click-through mechanisms | Medium | We expect this within core system, but may have to undertake development or use other tools |
| **4.7** | **Interaction and Web 2.0** | | | |
| 4.7.1 | Integration withTwitter, Delicious, Facebook etc | Ability to reference content using Web 2.0 technologies | Medium | Nice to have if available within core system, but may require development or another tool |
| 4.7.2 | Support for deep linking | Ability to create a link to page level | High | We expect this within core system |
| 4.7.3 | Personalisation of search | Ability to customise and save searches | Medium | Nice to have if available within core system |
| 4.7.4 | Personalisation of display | Ability to customise and save a presentation layout | Medium | Nice to have if available within core system |
| 4.7.5 | Personal user annotation | Ability for users to annotate digital objects, including marking sections of images | Medium | Nice to have if available within core system, but may require development or another tool |
| 4.7.6 | Bookmarking | Ability for users to bookmark digital objects (within the system) for later access | Medium | Nice to have if available within core system |
| 4.7.7 | Collaborative discussion environments (wiki) | Ability for users to provide public commentary on the digital object | Medium | Nice to have if available within core system, but may require development or another tool |
| 4.7.8 | Collaborative transcription tools | Ability for users to contribute transcriptions, tidy up the OCR etc | Medium | Nice to have if available within core system, but may require development or another tool |
| 4.7.9 | Public tagging | Ability for users to add public tags (as with Flickr) | Medium | Nice to have if available within core system, but may require development or another tool |

| | | | | |
|---|---|---|---|---|
| 4.7.10 | Public annotation | Ability for users to create public annotations | Medium | Nice to have if available within core system, but may require development or another tool |
| 4.7.11 | Text mining tools | Tools to support text mining | Low | Likely to require separate tools |
| 4.7.12 | Visualisation tools | Tools to support visualisations of content/collections | Low | Likely to require separate tools |
| 4.7.13 | Translation tools | Automatic translation tools | Low | Likely to require separate tools |
| **4.8** | **Commercialisation** | | | |
| 4.8.1 | Online ordering of reproductions | Ability for users to request high-quality copies | High | Likely to be provided by other tools or development work rather than core system |
| 4.8.2 | Online licensing of reproductions | Ability for users to request licenses for publication of content | Medium | Likely to be provided by other tools or development work rather than core system |
| 4.8.3 | Processing orders | Management of orders | High | Likely to be provided by other tools or development work rather than core system |
| 4.8.4 | Online payment | Ability for users to pay online using a credit card, PayPal or similar | High | Likely to be provided by other tools or development work rather than core system |
| 4.8.5 | Delivery of orders | Tools to support online delivery/download of purchased content | Medium | Likely to be provided by other tools or development work rather than core system |
| **4.9** | **Management reporting** | | | |
| 4.9.1 | Generation of reports | System must enable generation of reports/statistics of usage | High | We expect this within core system |
| 4.9.2 | Google analytics | Possible to use Google analytics to examine logs | Medium | We expect this within core system |
| **5** | **Additional requirements/information for any commercial systems** | | | |

| | | | | |
|---|---|---|---|---|
| 5.9.1 | Company | Established company with good track record, strong links with relevant community, large user base | Info required | |
| 5.9.2 | Support system | Good helpdesk system, fast response, can be reached during regular working hours | Info required | |
| 5.9.3 | Contract | Meets requirements of university procurement | Info required | |
| 5.9.4 | Cost | Competitive, within project budget | Info required | |
| 5.9.5 | % of any profits fed into development | What % of any profits are fed into product development? | Info required | |
| 5.9.6 | Training | Availability of on-site training | Info required | |
| 5.9.7 | Documentation | Availability of good documentation in English for the latest version | Info required | |
| 5.9.8 | Online help | Availability of online help, preferably context sensitive | Info required | |
| 5.9.9 | User groups | Active user group with similar membership to Library | Info required | |
| | | | | |