

CURL Members' Meeting 16 March 2005

Updating Bodley's Vision: the Google Mass Digitisation Project Ronald Milne

In December 2004 the Bodleian signed an agreement to become part of Google Print. This seeks to gain access to the Deep Web and the content of knowledge repositories as part of Google's mission to 'organize the world's information and make it universally accessible and useful'. The other libraries involved in the project are Harvard, Michigan, New York Public and Stanford.

The agreement is for three years initially, but may be extended if both parties wish. During that time Google will digitise 1-1.5 million of the Bodleian's 19th century out-of-copyright monographs and non-newspaper serials. The period was selected by the Bodleian with EEBO and ECCO in mind and to minimise copyright issues. The material to be digitised will be selected by an Oxford-employed local project manager. Google can only accept English, French, German, Italian and Spanish character sets at present, but this may change during the project. Digitisation will take place at Oxford, beginning June/July 2005. Page turning will be done manually to minimise impact on the works.

Virtually all costs apart from the local project manager and selection of material will be met by Google. Google will not charge for access but generate income through advertising. Oxford's digitised works will be in the form of OCR'd text plus images in JPEG 2000 format, indexed for fast, easy access, printable through Google's Print Services. There will be a link to the Bodleian's online catalogue, and a mybodley@google.com, or similar site, to enable users to view 'the Bodleian collection'. Page-to-page and around-book navigation will be possible.

A controlled-environment depository is being built in which the print set will be housed to the BS 5454 standard. It is hoped that as part of the project each item will be given an RFID tag or barcode to enable high density storage and automated retrieval.

The Bodleian views this project as an extension of the ethos it has maintained since its foundation in 1602, that of enabling access to its stock for those with reasonable need (60% of the Bodleian's users are not members of Oxford University). The project fits well with the Bodleian's hybrid library development, the 'Oxford Digital Library', and is work the Bodleian would not have been able to carry out itself.

Google will provide Oxford and the other libraries in Google Print their own sets of the digitised works so there will be two sites for the digital files, which each partner can exploit non-exclusively, plus the original print works in existence. Oxford can sub-license the rights grants to other UK legal deposit libraries if they covenant individually with Google, so the project will also help to maximise the chance of long-term preservation of the works.

Points arising from discussion following the presentation:

- There are technical matters to be resolved in relation to Oxford's own digitised set being made freely available as Oxford must prevent large-scale harvesting of the material.
- Google's focus is on making material from Oxford and the other Google Print libraries accessible. The project does not include consideration of a union catalogue or single interface to all the digitised collections, nor will complete bibliographic records necessarily be created for items lacking them, though metadata will be generated automatically. Is the future of COPAC and similar services to be as gateways to digitised content – would Google and other making material available in this way provide the sophistication of access of COPAC?
- It is likely Google would wish to digitise further collections if the initial five-library pilot goes well. The measure of success is not certain but is likely to be related to number of

accesses as advertising will generate Google's return. It's unlikely the participating libraries will know who is using their material.

- Google have a policy of continual improvement of images. Though the purpose of the project is to provide access to content, Oxford would be interested in discussing digital curation with CURL and the DCS.