

## OCR Accuracy Assessment

### A joint offering of King's Digital Consultancy Services and Digital Divide Data

*Validating the accuracy of OCR (Optical Character Recognition) to support planning for archival digitization projects*

#### Market Need

Automated recognition technologies have been a key driver of large scale text digitization by lowering costs, increasing search capabilities and delivering higher volumes of content than could be conceived with manual transcription. Optical Character Recognition (OCR) is often treated as a panacea for cheap, effective searching of large text resources.

However, both creators and users understand the limitations of OCR and the difficulties in gaining a true sense of the performance or accuracy of a search system when dealing with the inherently inaccurate text provided by automated technologies.

So the question becomes: "Just how inaccurate is my OCR and what can I do about it?"

Without solid statistical data on OCR accuracy, project leaders cannot appropriately decide how to optimize their OCR process, what other tools to use and what level of effort and cost to expend on text correction. Nor can they evaluate vendors' promised level of search performance, as the underlying textual accuracy remains an estimate rather than a fact. Better data on OCR accuracy in text output would help projects to make fundamental decisions on project design, technology and content selection.

#### Offering

King's Digital Consultancy Services in partnership with Digital Divide Data is thrilled to offer a service to assess the true statistical accuracy of OCR for digitization projects in any European language.

Our consultative service will use these accuracy results to give you actionable information that you can use to select content, optimize your OCR process, improve your search performance, design your delivery systems and reduce the costs for your project.

Our process involves a unique method for algorithmic comparison of OCR text output and original text transcripts to determine OCR accuracy to a high degree of confidence.

*Our deliverables include:*

#### **An OCR accuracy statistical report, including:**

- A complete item-by-item report of the OCR accuracy achieved in relation to the original text.
- OCR accuracy expressed as the percentage of characters, words, significant words, names, place names and numeric content rendered correctly.
- OCR accuracy expressed graphically for statistical overviews of the text resources broken down by factors such as date range, publication title, language or publication type.

#### **Actionable conclusions from this report, including:**

- Analysis of the reasons for the OCR accuracy results. For example, the age, condition, imaging technology or type and state of the text in the original can all affect OCR accuracy – this statistical method will help to identify the most likely causes and improve content selection.
- Recommendations for the means by which certain features of OCR performance can be optimized. For instance, for a publication with a large proportion of person names, improving the dictionary of names would improve OCR performance. This report will identify publications that require special attention.
- Results that could enable the comparison of different OCR engines to help the selection of the right technology.
- Analysis that will aid the assessment of a vendors' promised search engine performance against the likely performance suggested by the OCR accuracy results.

King's Digital Consultancy Services (KDCS) provides research and consulting services. Specialising in the information and digital domains our services are used by cultural, heritage and information professionals along with corporate clients. This venture is led by its Director, Simon Tanner, and is based within the Centre for Computing in the Humanities (CCH) at King's College London.

KDCS clients include: the British Library, National Library of Ireland, UNESCO, Tate Modern, Oxford University, Cambridge University, the National Library of Scotland, Stanford University, Kew Gardens, the Royal Academy of Arts, Imperial War Museum, the House of Commons, ITN News, BBC and many other organizations.

Digital Divide Data (DDD) bridges the divide that separates young people from opportunity by providing disadvantaged youth in Cambodia and Laos with the education and training they need to deliver world-class, competitively priced IT services to global clients and acquire essential business management skills. Clients include Yale University, The Reader's Digest Association, Kaplan Test Prep, Tufts Perseus Project, University Research Group, InSTEDD and Harvard Business School.

DDD was founded with the idea that the world's poorest citizens can produce their own solutions to poverty in the new global economy if they have access to the knowledge, skills, and opportunities that power economic growth and lasting change around the world.