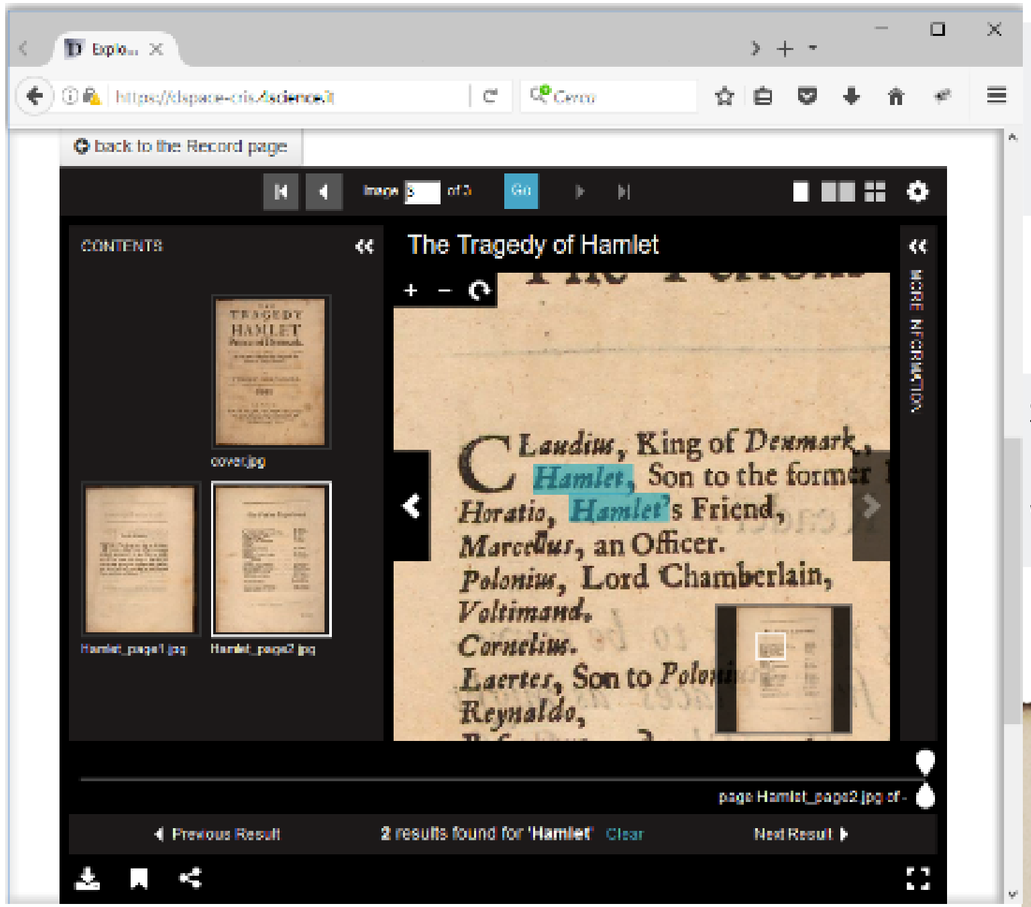
OCR&Transcription

www.4science.com

Get the text out of your digitized documents

For each image a curation task allows to extract its text representation in hOCR format for full-text indexing in SOLR. Tesseract supports a very large set of languages including: Italian, French, Spanish, German, Arabic, Simplified and Traditional Chinese and many others. The OCR engine can also be instructed with personalized training files to recognize fonts and specific languages.



DSpace integration with external Optical Character Recognition software;

Process hOCR format for full-text indexing in SOLR & image text overlay;

DSPACE Add-ons

Supports out-of-box a large set of languages, allowing more with personalized training files;

work online on manual transcription;

In the presence of the IIIF Image Viewer module, the OCR module also provides support for IIIF Search API through a server component, subject to the same terms of the module license. The IIIF Search API enable the activation of the search functionality inside the IIIF viewer, providing search within images, navigation through the results and highlighting on the image of the OCR text corresponding to the search terms entered.