ProQuest
Information and Learning

Digitisation of historical texts at ProQuest and ways of accessing variant word forms

Tristan Wilson,

ProQuest Information and Learning

www.proquest.co.uk

ProQuest
Information and Learning

Text Digitisation

- Databases include Early English Books Online (EEBO), Literature Online (LION), Historical Newspapers, British Periodicals...
- Digitisation methods: keying, OCR, conversion of texts supplied by partners
- Coding: SGML or XML, hierarchical

ProQuest
Information and Learning

Spelling Variations

- pleasant, pleasante, pleasannt, pleasent, pleasaunt, pleasaunte, plesand, plesant, plesaunt, pleafant...
- notwithstanding, notwithstandig, notwithstandinge, notwitstandinge, notwitstandiug, notwithwithstanding, notvvithstundinge, notwihstandynge, notwythstondynge...

ProQuest
Information and Learning

Browse / ‘Look For’ Lists

- Contain all words covered by a search
- Give user control

- Require user knowledge
- Can be laborious
- Not all variants located close to one another

ProQuest
Information and Learning

Typographical Variants Search Option

- User doesn’t need to know what to look for
- Tries every combination of expected variants

- Limited to typographical variants
- May retrieve irrelevant results

ProQuest
Information and Learning

Variant Forms Search Option

- Would specifically target known word forms
- User wouldn’t need to know what to look for
- More information could be supplied for advanced users

- Requires database
- Large numbers of variants might slow searches