

# Standards and Features

MSRI Digitization Workshop

16 April 2005

# Standards and Features Group:

- Ulf Rehmann
- David Ruddy
- Thomas Fischer
- Milena Dobрева
- Robert Miller
- Michael Doob
- Kevin Clancey, Recorder
- David Morrison, Chair

- Group will present recommended minimum standards and features
- We've made a start, but more work remains to be done
- This set of minimum standards/features should be accompanied by other “best practices” documents such as the 2003 report on Technical Standards from the DML planning group

# Recommended Minimum Standards and Features: Preface

- Grants may be given to existing projects to bring them up to standard level, for example, purchase of OCR software.
- MSRI should establish a standing committee to advise on compliance with these recommendations

# Recommended Minimum Standards and Features: I

- Retain original scanned images (in a lossless open format) at minimum 600 dpi.
- Recommend cover to cover scanning.

# Recommended Minimum Standards and Features: II

- Image to be accompanied by (OCR generated) text, keyed to the image.
- The OCR should be highly accurate (except for formulae) and sensitive to multilingual characters and isolated mathematical symbols
- [Open question: what standards should be used for text/image coordination?]

# Recommended Minimum Standards and Features: III (metadata)

- Must be shippable to MR/ZBL (in XML or BibTeX formats as they have specified).
- Some metadata should be published via OAI (using minidml at a minimum: format to be fully specified by our successors).
- There must be a user-accessible linked reference list for each article.

# Recommended Minimum Standards and Features: IV

- Each digitization project should be searchable, or should point users to a search engine



# Recommended Minimum Standards and Features: V

- Among delivery options it should be possible to obtain the entire article as a single file. (Books left for future discussion)
- Required delivery format for the present: PDF. Recommend that digitization projects also consider DjVu as a delivery format.
- Delivery formats should include the text layer.

# Recommended Minimum Standards and Features: VI

- Network addresses/item identifiers must be persistent (but the committee could not settle on a recommended method to guarantee this).