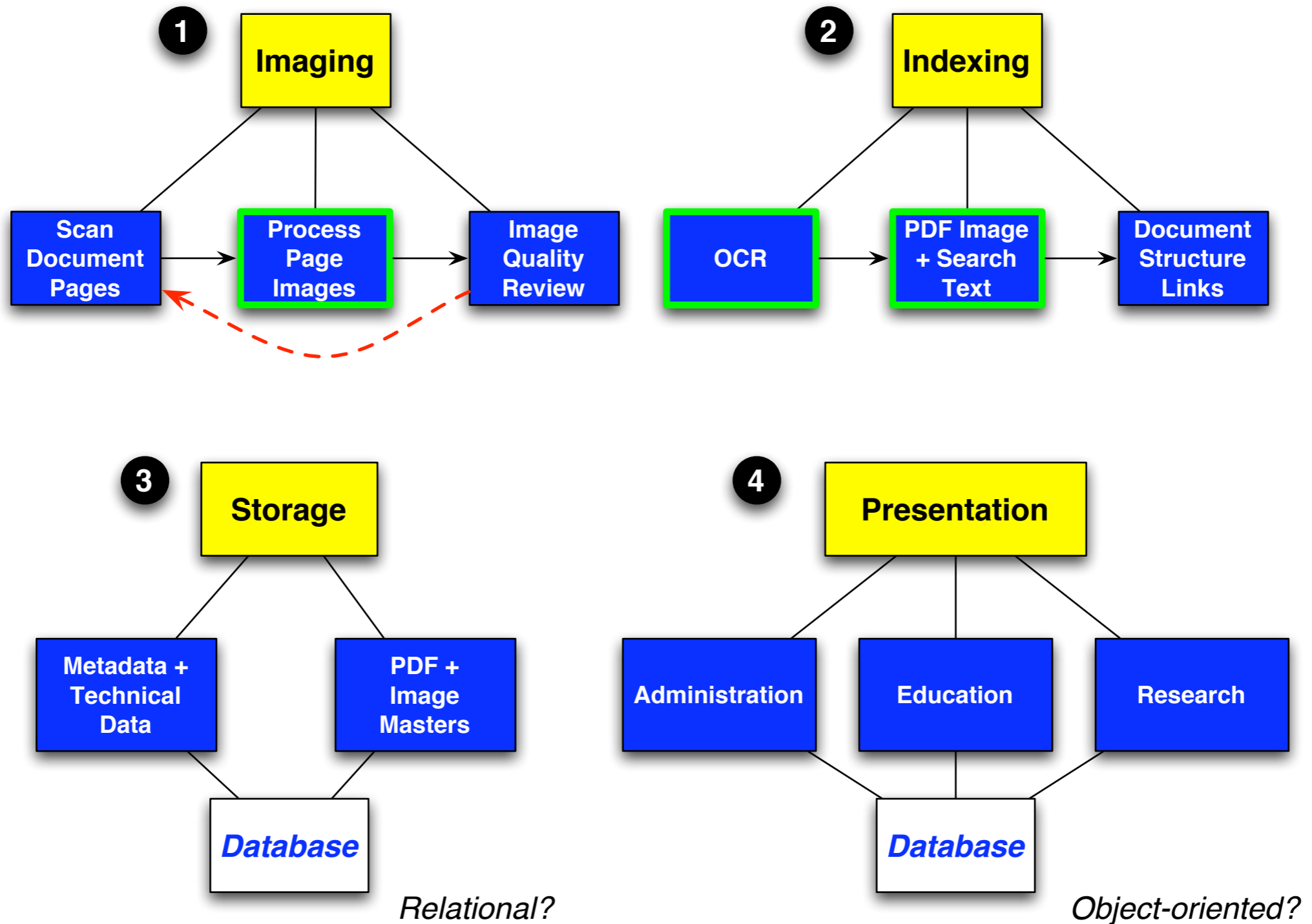

AMEEL

Looking Toward the Future

November 24, 2008

Digital Library Creation: Generalized Workflow



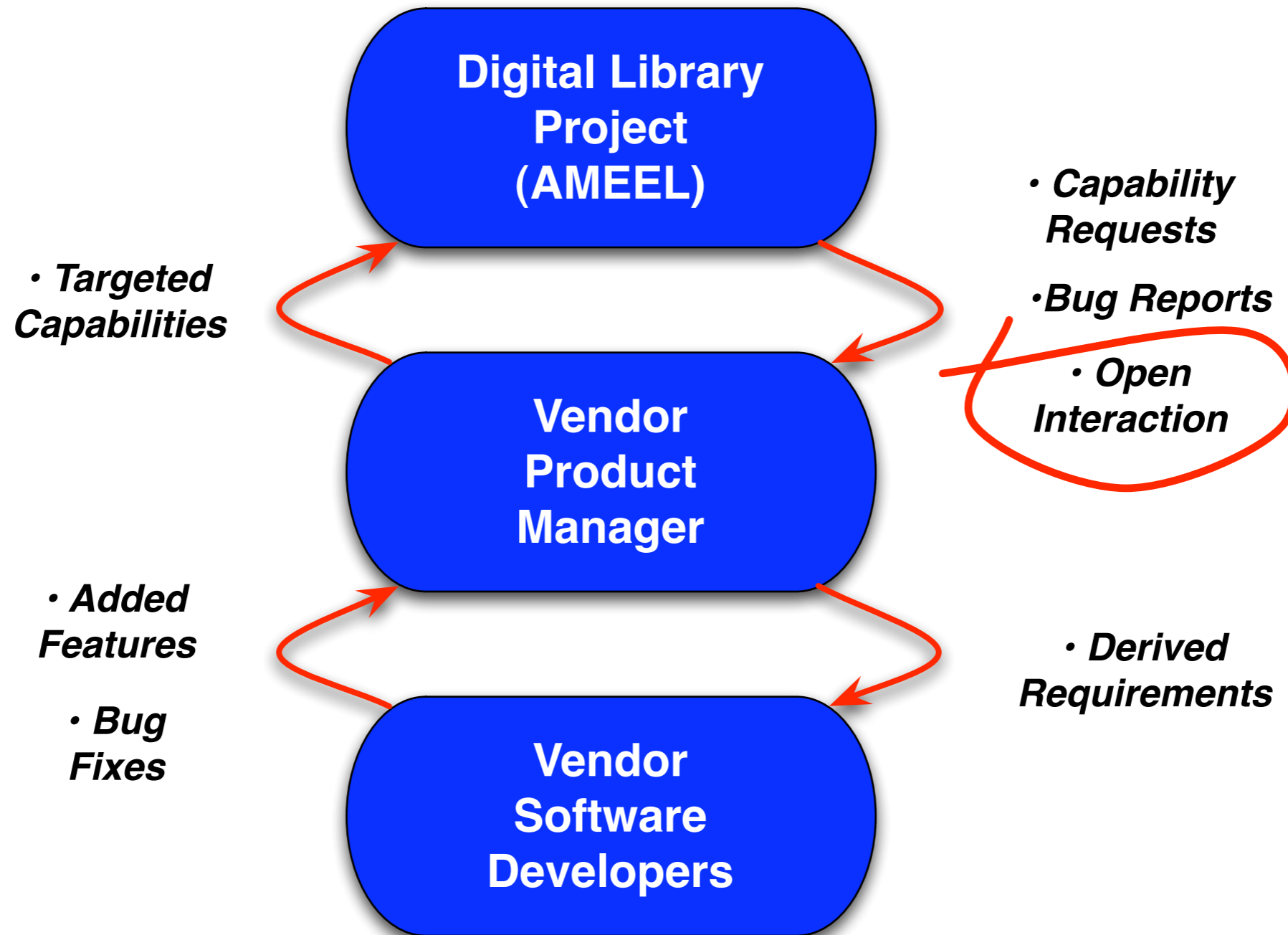
What are the real requirements?

Competing Priorities for OCR Component

User Needs	Government	Individual	Digital Library
Setup Specificity	Low	High	Very high
Degraded Source Materials	High	Low	Mixed
OCR Accuracy Requirement	Moderate	High	Very high
Repository Size	Extremely large	Extremely small	Moderate
OCR Interface	API only	GUI only	GUI and API
Search Support	Yes—Text	Maybe—Text	PDF Searchable Image
Machine Translation	Yes	Rare	No
Manual Intervention	Not acceptable	No problem!	Only as necessary

Can digital libraries live with a single (common) workflow?

Concrete Steps Toward Long-Term Goals



Focused public-private partnerships are worth considering in some cases

Knowing the Score—It's Important!

- Various OCR scoring methods are in use...
 - ◆ Count number of correct characters
 - ◆ Count number of correct words
 - ◆ And then there is the right way...
- The gold standard...UNLV scoring
 - ◆ Accounts for all error types: insertion, substitution and deletion
 - ◆ Handles UNICODE (multiple languages, embedded languages)
 - ◆ Requires groundtruth as input (double-keyed?)
 - ◆ Contact: Prof. Tom Nartker, UNLV ISRI (tom@cs.unlv.edu)
- Check definitions before comparing accuracy rates!

Enhancing Search

- Visually display query hits on page images
 - ◆ OCR must maintain image-text correspondence
 - ◆ Most modern OCRs maintain lots of geometric information
- Move beyond simple keywords
 - ◆ Boolean search
 - ◆ Lucene query language
 - ◆ Morphological analysis (e.g., AppTek, Business Objects[InXight])
 - ◆ Fuzzy matching?
 - ◆ Enhance information content of “query hit” scores
 - + Integrate OCR & information retrieval (IR) metrics

Degraded Images

□ Sources

- ◆ Copier, fax, old printer
- ◆ Aging process—yellowed
- ◆ Newspaper print (font, print issues; bleed through)
- ◆ Water damage
- ◆ plus much more!

□ Remedy?

- ◆ Image processing
- ◆ Automation often possible
- ◆ Interactive filter selection for stubborn cases