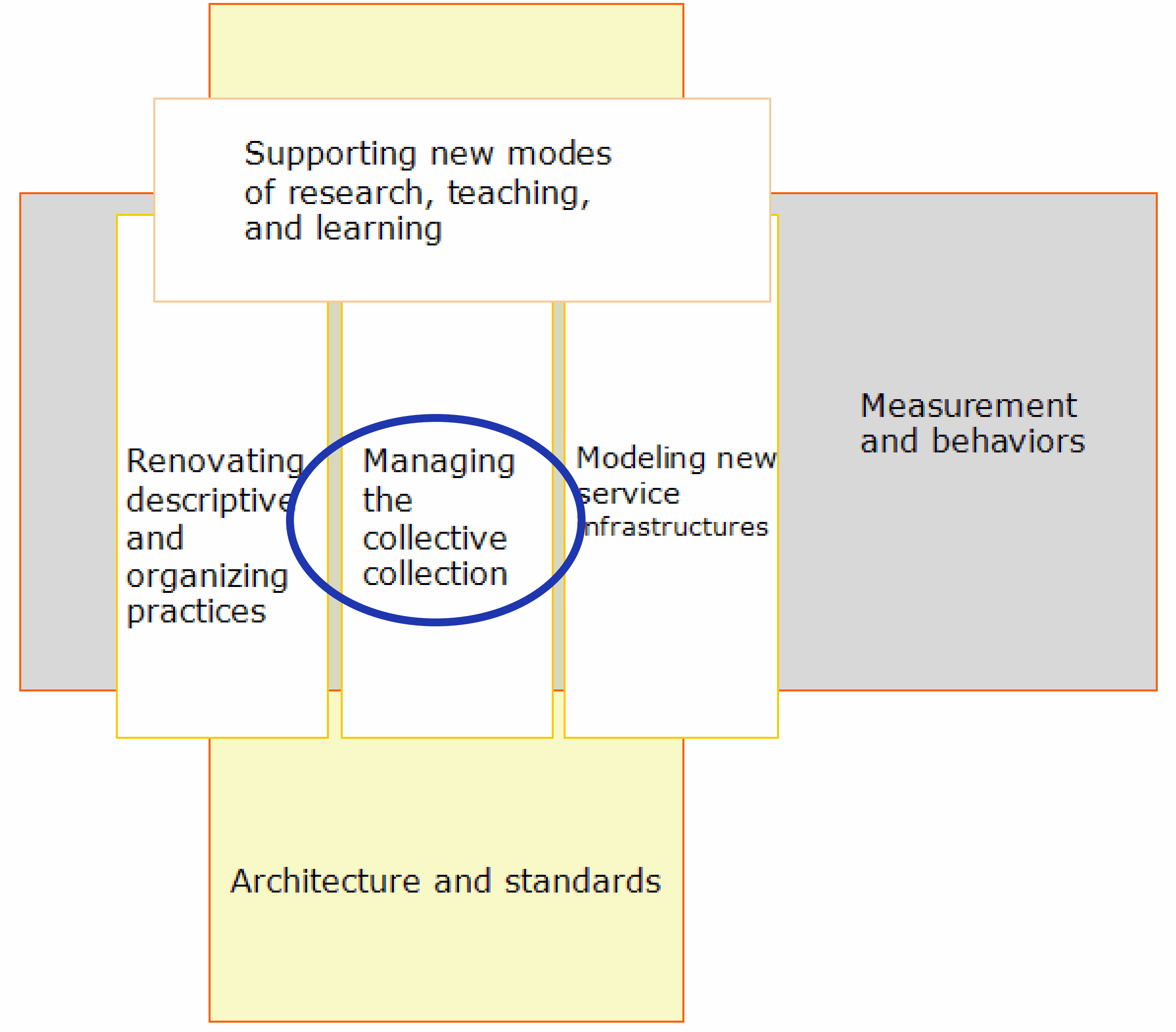
Mass Digitization and the Collective Collection

Constance Malpas
DLF Fall Forum 2006
Boston, MA

10 November 2006

malpasc@oclc.org

OCLC Programs & Research



Supporting new modes of research, teaching, and learning

Renovating descriptive and organizing practices

Managing the collective collection

Modeling new service infrastructures

Measurement and behaviors

Architecture and standards

Defining the Collective Collection

- System-wide print collection
  - “Anatomy of Aggregate Collections” (Dempsey et al, 2005)
  - “Books without Boundaries” (Lavoie & Schonfeld, 2006)
- Mass digitization and the ‘universal library’
  - “Scan this Book” (Kelly, 2006)
  - “*Quand Google défie l'Europe*” (Jeanneney, 2005)
- Optimizing network resources
  - Janus conference: “recon” challenge (Atkinson, 2005)
  - Digitisation in the UK - Loughborough Study (CURL/JISC, 2005)

Distributed Corpus

- Google Book Search 15M titles by 2010
- IA / Open Content Alliance 37K+ titles 11/2006

American Libraries	**4,900 titles (and growing)**
Canadian Libraries	**4,700 titles (and growing)**
Million Book Project	**10,000 titles**
Open Source Books	**1,500 items**
Project Gutenberg	**7,000 items**

- European Digital Library 6M titles by 2010
- AlouetteCanada 1M items by 2007
- Legacy and licensed collections

Are these books without boundaries?

Survey on library partnerships

- 12 “hard questions”
- Challenge: identify 3 most pressing concerns
- Opportunity: identify (and fill) gaps

- Distributed to all DLF registrants
- 61 responses as of 6 November 2006
- Represents US (88%); Canada, UK, EU (12%)

- In general:
  - There is interest in all of the questions we posed
  - The really important questions really stand out
  - 17% of respondents supplied an additional question

Most Urgent Questions

- Is a **regional, national, or multi-national framework for digitization** desirable or feasible? What would you expect such a framework to contribute?
- Who is responsible for ensuring the **persistence of the aggregate collections to which you have contributed**?
- What does your institution know (or need to know) about **how users are interacting with the outputs of mass digitization**?
- As a contributor to the mass digitization enterprise, do you expect other institutions or other entities to be able to **build collections or services on the materials you've contributed?**
- What is the **most significant compromise your institution has made to enable participation** in mass digitization?

And now for the answers …

Robin Chandler, Director of Data Acquisitions, California Digital Library

Open Content Alliance + Google Book Search

Sian Meikle, Digital Services Librarian, University of Toronto

Open Content Alliance + AlouetteCanada

Barbara Taranto, Director of the Digital Library Program, New York Public Library

Google Book Search + large-scale legacy collection

Stuart Dempster, Manager of Digitisation Programmes, UK Joint Information Systems Committee

Network coordination of national digitization efforts

Urgent Question (1)

- Is a **regional, national, or multi-national framework for digitization** desirable or feasible? What would you expect such a framework to contribute?

Related user-contributed question:

- How might libraries within and outside of mass digitization projects work together to build as complete as possible retrospective digital library?

Urgent Question (2)

- Who is responsible for ensuring the **persistence of the aggregate collections to which you have contributed**?

Related user-contributed question:

- How are the “root” (library-contributed) collections being preserved, and by whom?

Urgent Question (3)

- What does your institution know (or need to know) about **how users are interacting with the outputs of mass digitization**?

Related user-contributed question:

- Given that some of the same commercial interests that are financing mass digitization in libraries have a business model based on directing advertising to Internet users related to what types of information they're seeking, do you feel there is *reason to be concerned at all about potential breaches of confidentiality of library patrons* who are interacting with the outputs of mass digitization?

Urgent Question (4)

- As a contributor to the mass digitization enterprise, do you expect other institutions or other entities to be able to **build collections or services on the materials you've contributed?**

Related user-contributed question:

- Do we imagine *a gigantic corpus of text* (and images and sound), to be analyzed and explored as a unity? Or a collection of *bibliographic/visual/sonic "atoms,"* each discrete and separately discoverable? Or *intermediate aggregations* built around particular subjects or other concerns?

Urgent Question (5)

- What is the **most significant compromise your institution has made to enable participation** in mass digitization?

Related user-contributed question:

- Has your *participation in mass digitization efforts required you to compromise or limit any piece of the usual service* you provide to digitized materials, and if so, how did you justify it?

Community Concerns: Beyond Books

- The most widely-publicized mass digitization efforts are focused on conversion of printed books. How do you justify the **community investment in improving access to works that are (relatively) widely held**?

Related user-contributed questions:

- What is the potential for "format-blind" digital aggregations?
- What is the place of special collections in mass digitization?
- What about legacy collections?

Community Concerns: Rights, rights, rights

Rights of content creators:

- What role did *copyright considerations play in your discussions* prior to the decision to participate in mass digitization? Do you feel things in that area are playing out as you expected?
- Copyright/access rights are one of the biggest challenges facing institutions considering mass digitization: what could be possible to *avoid the 20th century black hole*?

Rights of content providers:

- How can institutions collaborate to identify non-commercial sources of funding for mass digitization efforts to insure true *open access to materials* that are being digitized?

Rights of readers and researchers:

- Given that some of the same commercial interests that are financing mass digitization in libraries have a business model based on directing advertising to Internet users related to what types of information they're seeking, do you feel there is *reason to be concerned at all about potential breaches of confidentiality of library patrons* who are interacting with the outputs of mass digitization?