New Paradigms for Knowledge Sharing in the Scholarly World:
Developing Institutional Repositories

Tyler O. Walters
Associate Director for Technology & Resource Services
Georgia Tech Library and Information Center

Consortium of Southern Biomedical Libraries
March 11, 2006
Augusta, GA
Today’s Afternoon Agenda

- A Bit of History and the Overall Paradigm
  - Scholarly Communications / the Information Life Cycle / “the call” for IRs

- Trends and Developments
  - Plus a technology overview – software, software models, and partnerships to make it happen

- Development of the GT IR, SMARTech

- Things the GT Library is thinking about today
  - Value added services and functionalities
  - New modules
  - Software integration
  - Preservation networking
Today’s Afternoon Agenda

Group project:

- **Mission and Vision:**
  What should your IR do for your community? What will you collect? What are the value-adds?

- **Marketing Plan:**
  Getting “buy-in” and comprehending the service. Critique of the SMARTech marketing plan, applying it to your institution.
A little about my background:

- M.A. in Archival Management (NCSU)
- M.A. LIS (Univ. of Arizona)
  - Subjects: labor/social history, urban anthropology/sociology, knowledge mgmt., org. behavior, sociology of information
- GT Library Digital Initiatives, 2002-present
- IPST Digital Library Program, 1998-2002
- Iowa State Univ., Special Collections, 1992-98
  - Web site, digital collections, first ETD ’96, e-records, metadata
- Northwestern University Library, 1988-2002 gopher server
- North Carolina Div. of Archives & History, 1986-88
  - Online cataloging photos, thesaurus building
the overall paradigm of...

Scholarly Communications

...is the system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use. The system includes both formal means of communication, such as publication in peer-reviewed journals, and informal channels such as electronic listservs.  

http://osc.universityofcalifornia.edu/
Types of Scholarly Communication

- **Formal, i.e.**
  - Journal publications, research papers, technical reports, working papers, conference papers, lectures, records, personal papers

- **Informal, i.e.**
  - Listservs, threaded discussion lists, virtual community sites / collaboration spaces, blogs, wikis, e-mail, etc.
Sources of Information Production

- **External information**
  (information objects that come to us from beyond our organization – i.e. publishers, commercial content providers)
  
  - journal publications, post-prints, conference proceedings, text books, commercially-packaged instructional materials, wikis/blogs, e-mail

- **Internal information**
  (information objects that come to us from inside our organization – i.e. university units, faculty)
  
  - research reports, research data sets, learning objects, faculty’s instructional materials, white papers, lectures and symposia, wikis/blogs, e-mail, institution’s records, personal papers
The Life-Cycle of Information

Production (internal info & external info)

Long-term sustainability (preservation)

Information Use & Integration

Collection/Organization/Access (IRs)
Early Calls for Institutional Repositories (IRs)

- **Budapest Open Access Initiative (BOAI)**
  
  [http://www.soros.org/openaccess/](http://www.soros.org/openaccess/)
  
  Feb. 2002 - recommends two strategies:

  1. self-archiving
  2. open access journals.

- **Scholarly Publishing and Academic Resources Coalition (SPARC)**
  
  [http://www.arl.org/sparc/](http://www.arl.org/sparc/)
  
  July 2002 - “the case for inst. repositories”

  - promotes both open access journals and the development of institutional repositories. SPARC has a number of open access partners
Institutional Repositories

Role of Institutional Repositories in Scholarly Communications:

• **What are IRs?**
  
  “a digital archive of the intellectual product created by the faculty, research staff, and students of an institution and accessible to end users both within and outside of the institution, with few if any barriers to access. The content… is institutionally defined, scholarly, cumulative and perpetual, and open and interoperable.” (Rick Johnson, SPARC, 2002)

• **Collect / Organize / Distribute content**
Repository Concepts

- **Disciplinary repositories**
  facilitates sharing and storage of discipline-based research materials. These repositories ("e-print servers") have high rates of participation in their respective fields. Repositories exist in disciplines such as classical literature, history of philosophy, economics, chemistry, cognitive sciences, mathematics, and physics.

- **Institutional / Organizational repositories**
  digital archives of intellectual products created by the faculty, staff, and students of an institution and accessible to end users both within and without the institution, with few if any barriers to access.

- **Collection-based repositories**
  support finding, producing, and curating the information of a group or community within an institution.

- **Self-archiving**
  depositing by the author of a digital document in a publicly accessible institutional or disciplinary repository via a web site. Includes articles and preprints by individual researchers.
Approaches:
Discipline & organization-wide

Discipline Model:

- Pre-print Archives
  Physics, math, computer sciences, quantitative biology

Organization-wide Strategies and Models:

- University of Kansas Inst.-wide digital preservation
  [https://kuscholarworks.ku.edu/dspace/handle/1808/166](https://kuscholarworks.ku.edu/dspace/handle/1808/166)

- NARA Electronic Records Archive
Approaches: Collection-based

  CDL – [http://dermatology.cdlib.org](http://dermatology.cdlib.org)

- Conference Proceedings

- Electronic Theses and Dissertations

- Student E-Portfolios
  Duke – [https://portfolio.oit.duke.edu/index.jsp](https://portfolio.oit.duke.edu/index.jsp)

- Data Sets
Other IR Uses

- Digital Object Productivity Repository
- Records Management
- Faculty / Student Portfolios
- Learning Objects and Course Captures
Some Trends

- **IR Services**
  Includes client services like digitizing/scanning, metadata services, etc.

- **Self-Archiving vs. Library Archiving**
  “Library collection development” approach or self-selection

- **Repository Certification**
  Foundation of an institution’s plan for information management

- **Data Curation /Management**
  Collaborative opportunities for libraries and the sciences
Some Trends

- **IR services**: Cornell Univ. Library
  http://dcaps.library.cornell.edu/
  http://www.library.cornell.edu/iris/migration

- **Self archiving vs. Library archiving**: SHERPA Project
  http://www.sherpa.ac.uk/romeo.php
  http://romeo.eprints.org/publishers.html

- **Rep. analysis, certification, “trusted repositories”**
  JHU / Mellon Fdn.  http://ldp.library.jhu.edu/projects/repository
  CRL:  http://wwwcrl.uchicago.edu/content.asp?l1=13&l2=58&l3=142

- **Data curation/management** - UK Digital Curation Centre
  http://www.dcc.ac.uk
Discussion:
SC/IR movement, trends, developments

- Positives?
- Negatives?
- Interesting?
IRs :: Technology Overview

- Repository software
  - “Choosing software for an IR,” Jody DeRidder, 2004
    - [http://diglib.lib.utk.edu/dlc/ir_software.pdf](http://diglib.lib.utk.edu/dlc/ir_software.pdf)

- Many repository software systems are available:
  - DSpace ([http://www.dspace.org](http://www.dspace.org))
  - Fedora ([http://www.fedora.info/](http://www.fedora.info/))
  - Digitool ([http://www.exlibris.co.il/digitool.htm](http://www.exlibris.co.il/digitool.htm))
  - BEPress ([http://www.bepress.com](http://www.bepress.com))
  - E-prints ([http://www.eprints.org](http://www.eprints.org))
Software Models

- **Commercial software**
  (company support, company development, possible rising costs)
  - Digitool, ContentDM, Harvest Road

- **Application service providers**
  (hosted on another’s server, provide software, storage, etc.)
  - BEPress

- **Open source software**
  (community support, local development, possible lessening costs)
  - DSpace, ePrints, Fedora
Partnerships to “Make it Happen”

- Main Library (GT)
  http://SMARTech.gatech.edu

- w/ Campus IT Division (Ohio State U)
  https://kb.osu.edu/dspace/index.jsp

- Statewide (CARL, Galileo)
  http://www.coalliance.org/Homedocs/IR_TF.html

- Regionally / College Affiliation (ACS)
  http://spumoni.colleges.org/elated/about.jsp?jsessionid=602135922CD6B95F2B14FCCE6D45550C
Discussion: technology overview

- Positives?
- Negatives?
- Interesting?
What is SMARTech?

- SMARTech, or Scholarly Materials And Research @ Georgia Tech, is a repository for the capture of the intellectual output of the Institute in support of its teaching and research missions. SMARTech connects stockpiles of digital materials currently in existence throughout campus to create a cohesive, useful, sustainable repository available to Georgia Tech and the world.”

- DSpace on Sun Solaris / UNIX (move to Linux)
- 6,000 objects (03/06, doubled in 17 months)
- ca. 5th largest DSpace installation, 40+ (sub)communities, 70+ collections
- 190,000 object views / 21,000 searches (past 10 months)
Welcome to SMARTech

Search:
Enter text in the box below to search the collections.

Advanced Search

Browse Communities in SMARTech:
- Archives and Records Management [344]
- Auxiliary Services [12]
- Center for Experimental Research in Computer Systems (CERCIS) [61]
- Center for the Enhancement of Teaching and Learning (CETL) [11]
- College of Architecture (CoA) [48]
- College of Computing (CoC) [299]
- College of Engineering (CoE) [919]
- College of Liberal Arts (IAL) [35]
- College of Management (CoM) [22]
- College of Sciences (CoS) [159]
- Georgia Tech Student Publications [3]
- Georgia Tech Theses and Dissertations [1479]
- Graphics, Visualization, and Usability Center (GVU Center) [485]
- Institute Communications and Public Affairs (ICPA) [286]
- Institute of Paper Science and Technology (IPST) [3682]
- Institutional Research and Planning (IRP) [6]
- Library and Information Center [65]
- President's Scholarship Program (PS Program) [8]
- WERK [7]

If you would like to know more about SMARTech or would like to become an adopter, contact Katherine Jamieson.
The Digital Initiatives Department is…

- charged with building effective, dynamic knowledge & research management systems to preserve and provide access to the intellectual output of Georgia Tech. Its responsibilities include identifying, assessing, collecting, preserving, providing access, and making this output more valuable through digital information technologies, whether “born-digital” or convertible to digital formats. The Dept. provides these resources and services in support of the research and educational endeavors of the Georgia Tech community and to scholars around the world.”
SMARTech: Content

- GT-produced intellectual output from educational & research programs, communications activities:
  - Annual Reports
  - Computer Programs
  - Conference Papers
  - Data Sets
  - Learning/Complex Objects (“captured” courses [i.e. digitally recorded], multimedia simulations / visualizations (cognitive tools), textual documents, captured notes from faculty and students
  - Models
  - Lecture series materials
  - Pre-Prints/Post-Prints
  - Proceedings
  - Research Reports
  - Simulations
  - Technical Reports & Working Papers
  - Web Pages
  - White papers
SMARTech: Policy Issues

- Purposely refrained from devising many abstract policies
  - don’t want to inhibit growth and development of SMARTech

- Allow campus community to suggest SMARTech contents
  - One Requirement: GT-produced digital intellectual output

- Acceptable digital file formats:
  - Standard formats -- commit to migrate & access over long term
  - A variety of text, image, audio, video, and authoring software files

- Copyright: follow existing guidelines. If problem, take down

- Collection Development: follow library/arch. policies (’04)

- Devised our own Collecting Policy (’06)
SMARTech: Context & Planning

- Organizational commitment to Digital Initiatives in Library strategic plan, 2002:

- A major objective: “digital publishing program to promote open standards for faculty and students, resulting in durable, functional digital works accessible worldwide.”

SMARTech: Planning 2003

- Completed TRS Div. & DI Dept. strategic plans:
  - IR based on information needs of users
  - Systems/policy: capture, distribution, access, preservation
  - Campus digital initiatives ➔ promote open source approaches
  - Capture “at-risk” digital intellectual output (?)
  - Create digital resources from existing library resources
  - Staffing “moves” in Digital Initiatives and Systems Departments...
SMARTech: 2003 Drivers

- Technology development, born-digital prevalence concern: losing unit web site information not “sent” to Library
- Lib. budget growth: $1 mill (fy04), $750,000 (fy03) vote of confidence, Library can deliver on managing digital information
- GT mission: knowledge creation, transfer, and the library’s role
- Interdisciplinary research & cross-disciplinary data mining (Digital Repository, YES and “Silos” - NO)
- Acceptance of in ETD @ GT program (Fall 2003)
- Growth in web-based papers, reports, lectures (50+ examples, 2003)
Promoting Library services in a digital context

- Building upon academic departments disparate attempts to share information through...
  
  - Standard metadata to discover resources inter-institutionally
  - Digital preservation schemes and approaches
  - Online access over time, across technologies (open source)
  - Searching/knowledge mining (cross-disciplinary discovery)
SMARTech: “Fast-tracking”

- **Why?**
  Because despite users telling us they would like to have an IR, they, and we, don’t really know how they will interact with it.

- **Response to this dilemma?**
  We built a “pilot” IR asap to observe users’ interaction with it and gain their feedback re: their needs.

- **Pressure to fast-track?**
  *You bet.* We need to produce and show the value of the Digital Initiatives program. With $1.75 million added to library budget in 2 years, we need to show ROI to our resource allocators.
SMARTech: Software Selection

- Formed library team to review, assess, select IR software, and review “functional issues”:
  - Preservation/admin. metadata handling
  - Subject headings / controlled vocabulary, and end user utility or need
  - Submission process (user friendliness)
  - Policy issues, i.e. who has rights regarding access, submission, review, etc.
  - Searching, basic vs. advanced
  - Fiscal costs to install, maintain
  - Open source as a mission, a driver
  - Long-term preservation?
SMARTech: Software & Launch

- Reviewed DSpace and EPrints more closely
  - GT ECE / Library: DSpace, learning object mgmt. (03/04)
  - Pilot implementation, Spring 2004
  - Public launch of SMARTech, August 1, 2004
  - Over 3,000 objects (10/04) / 10,000 hits in first 2 months
  - Started with 8 communities, 11 collections

- Why DSpace for GT?
  - Momentum – university consortial support
  - Functionally, best met our basic needs
  - Open source, met objectives to be repository development leader
SMARTech

- Overall Objective:

  • Not to achieve perfection, but rather to start down the “IR road.” Adjust along the way.
  • Our goals must be obtainable within the ability, energy, and time available from our programmers, staff, and managers.
Discussion

- Questions / comments on GT’s history of development?
- Lessons learned thus far?
- Positive? / Negative? / Interesting?
SMARTech

- So, where are we today?
  - Marketing / Collecting (library vs. self-submission)
  - Library - Organizational Maturity
  - “Value-added” services developed for faculty, “interweaving IR into GT “information fabric”

Digital Preservation:
MetaArchive Preservation Network
ASERL LOCKSS - ETD Initiative
SMARTech: Marketing Plan

- Situation analysis
  - Marketing advantages
  - Marketing challenges

- Target audience
  - Faculty/Researchers
  - Students

- Goals
  - Measurable benchmarks

- Strategies and tactics
SMARTech: Collecting Plan

- Purpose of SMARTech
- Collecting priorities: Subject areas and forms of intellectual output
  - “Centers of Excellence,” Emerging areas of significance
  - Materials of specific interest, Electronic publications
- Assessment Criteria
  - Considerations: content, technical, legal
- Digital File Formats
  - Supported, known, unsupported
- Deposit procedures
- Permanance/Deaccessioning/Suppression/Removal
- Access / Preservation
Organizational “System:” Managing digital intellectual output services

- Marketing
- Collecting
- User instruction/support
- IP / Copyright issues
- Technical project management
- Preservation
- Metadata
- Application mgmt./content production/workflows
- Application development/programming
- Systems administration
SMARTech: “The Future”

Future Projects: (’04)

- Faculty / student portfolios (Oct. ’04 ??)
- Distributed learning object repositories (Oct ’04??)

’06 Planning:

- Marketing, collecting, implementing value-adds, i.e. restriction capabilities, P/T document output, researcher pages, taking on dept-level databases of faculty output, etc. (Feb. ’06)
Services being developed in IRs / SMARTech:

- “My DigitalArchive:” Researcher pages
- Syndication of content to portals, dept. web sites
- Levels of Restricted access
  - Managing proprietary and embargoed information
- P/T document output
- Co-submission of sponsored research reports
- More software integration, WebCT, EndNote, UPortal
- Improved searching / federated searching
- Goggle Scholar / citation analysis
- Current projects: 1) Coll. of Architecture 2) School of Civil and Environmental Engineering
Digital Preservation

- Need strategies beyond files in a repository structure!

  - MetaArchive Preservation Network
    - http://metaarchive.org
    - http://smartech.gatech.edu:8282/dspace/handle/1853/7287

    - **Decentralized Approach** (question the “one copy at one institution” approach)

    - **Built on LOCKSS** (supports “distributed digital replication” approach). Peer-to-peer network architecture, each node communicates with each other to replicate content. Minimum of six nodes

    - **Dark Archive** (preservation with no public access. High accessibility = high costs)

    - **Automated format migration tools**

    - **Low Cost** (Planned minimal expense, low barriers to adoption for mid-size insts.)

    - **Flexible, adaptable multi-institutional model**

    - **LC / NDIIPP partnership…** (1 of 8 initial)
Discussion on future development

- Positives?
- Negatives?
- Interesting?
“Exercising!”

Group discussions:

- **Mission and Vision:**
  What do you want your IR to do for your community? Value-adds?

- **Marketing Plan:**
  Getting “buy-in” and comprehending the service. Critique the GT SMARTech plan, apply it to your institution.
Thank you!

Tyler O. Walters
Associate Director for Technology & Resource Services
Georgia Tech Library and Information Center

404-385-4489
Tyler@gatech.edu