Digital Preservation: From Projects to Infrastructure

Margaret Hedstrom
University of Michigan

Outline of the Presentation

- Recent Developments in Digital Preservation
- Current Approaches and Solutions
- Infrastructure Requirements
- Bridging the Gaps
- Conclusion

Digital Preservation Challenges

- Keeping information alive and accessible in spite of changing technology
- Ensuring that information is credible and understandable so that it is not used inappropriately
- Sustaining information with an adequate flow of revenue over many decades

Emerging Standards and Best Practices

- Framework and Models for Trusted Repositories
- Standards for Metadata and Data Formats
- Some Tools Managing Technology Dependencies

New Challenges

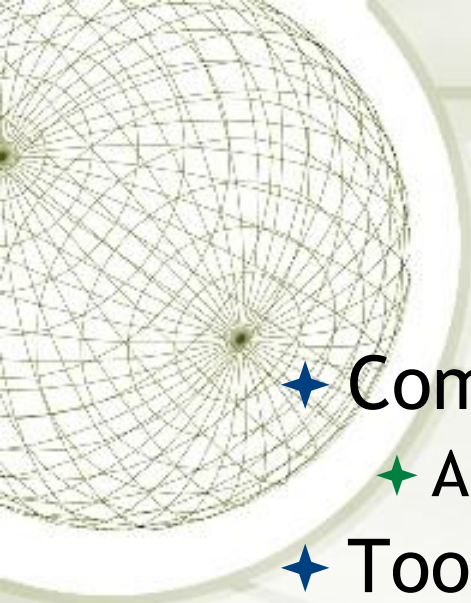
- Need for digital preservation repositories and services in new environments
  - Scientific Data
  - Entertainment and New Media
  - Personal Archives
- Need for interoperability across repositories
- Need for integration of data and publications

New Challenges

- Scalability of current methods
  - Diversity of data, formats, production environments
  - Quantity of ubiquitous data
  - Appraisal and Selection
  - Costs of digital preservation
- Need for approaches that generalize and scale gracefully

Moment of opportunity

- The pieces of a global network are falling into place
  - Computation
  - Communication
  - Content
- Or are they?
  - Diversity of content?
  - Content exploitation?
  - Comprehension?
  - New knowledge generation?

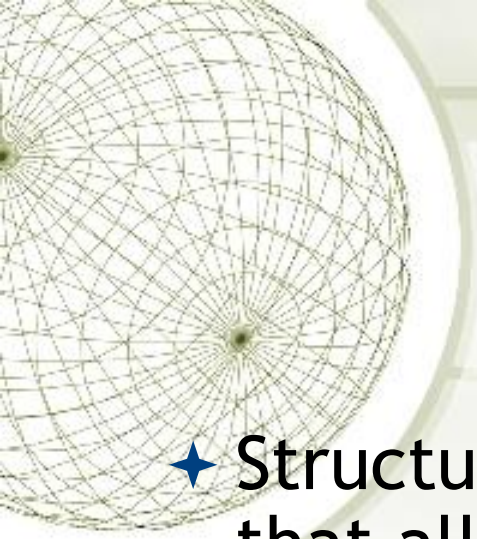


What is missing?

- Comprehensive content
  - Across disciplines, language, location
- Tools for analysis
- Sharing and exchange of content, data, results
- Acceleration in the generation of new knowledge
- Fundamental, not incremental, new discoveries
- Infrastructure to enable all of the above

Moving from Projects to Infrastructure

- Digital Preservation Projects have produced useful models, tools, and practices for specific types of content in specific environments
- How can we build on these projects and shift toward building digital preservation infrastructure?



What is infrastructure?

- Structures, systems, and social agreements that all allow disparate components of a system to work together on a grand scale.
- Effective infrastructure allows people to interact with systems easily.
- Useful infrastructure allows people to accomplish goals that would be impossible to achieve without it.

Digital Preservation Infrastructure Components

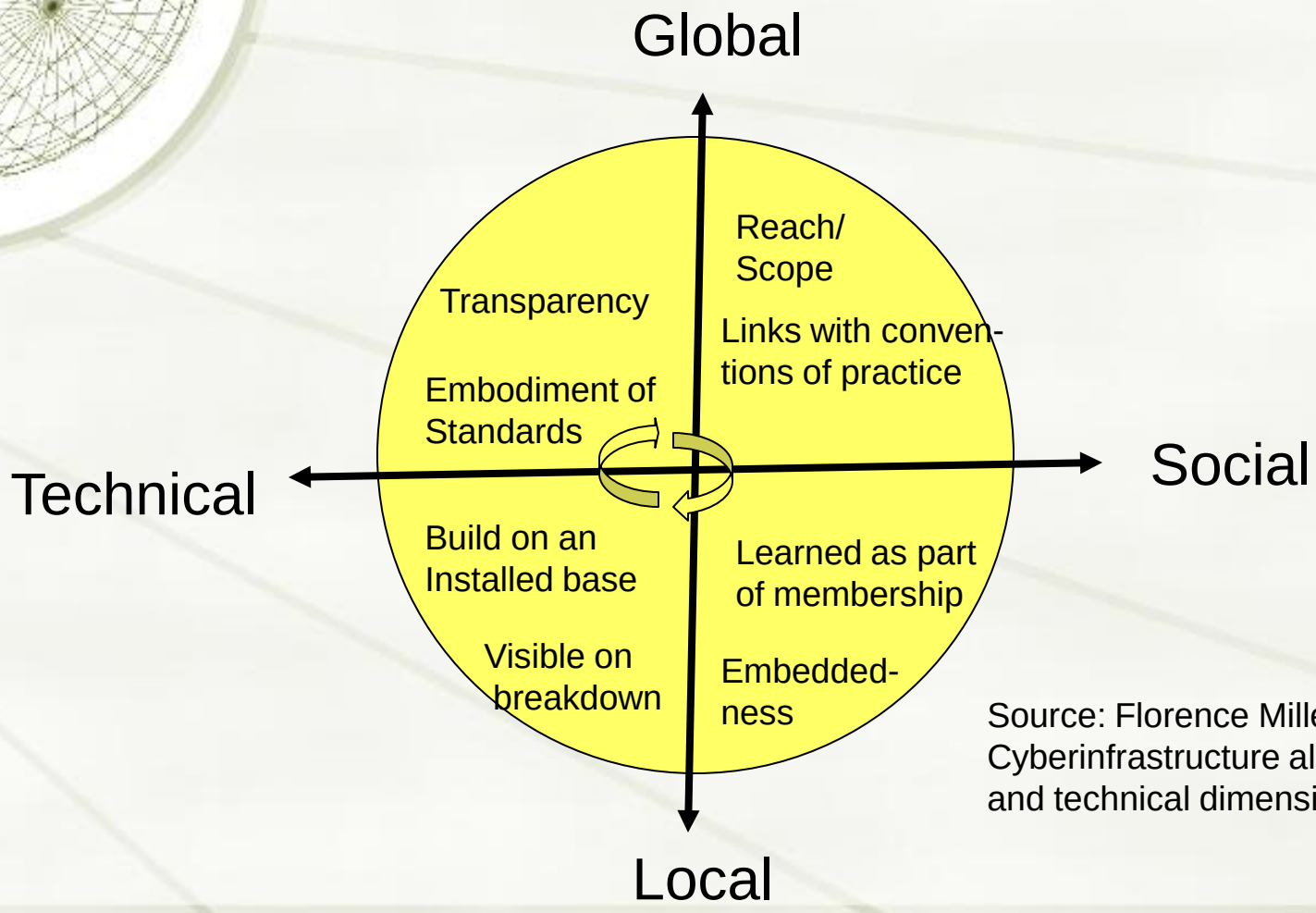
- Technical Aspects
  - Interoperable hardware, software, and networking components
- Intellectual Components
  - Interoperable metadata schema, ontologies, and knowledge representation
- Social Components
  - Agreement on roles and responsibilities, incentives and rewards

Characteristics of Infrastructure

- Embeddedness
- Transparency
- Reach or scope
- Linked with conventions of practice
- Embodiment of standards
- Built on an installed base
- Becomes visible upon breakdown
- Is fixed in modular increments, not all at once or globally

Karen Ruhleder and Susan Leigh Star

Infrastructure Requirements



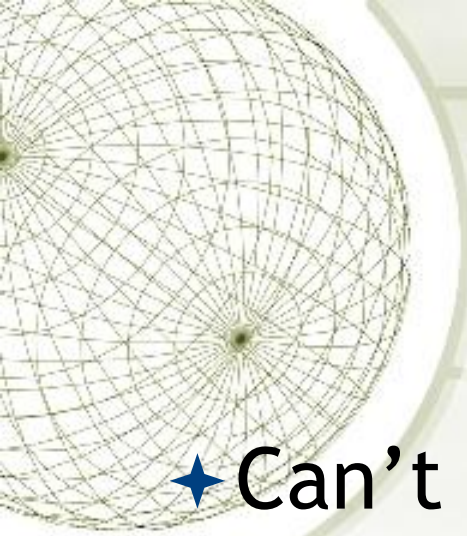
Source: Florence Millerand, Cyberinfrastructure along social and technical dimensions

Infrastrcture: Some Concrete Examples

- The power system
- The transportation system

Cyber-infrastructure Initiatives

- Digital Projects and Digital Libraries
- [US] National Science Foundation (NSF) Blue Ribbon Panel on Cyberinfrastructure for Science and Engineering
- E-Science and Information Society Initiatives
- ACLS Commission on Cyberinfrastructure for Humanities and Social Science
- CASPAR Project



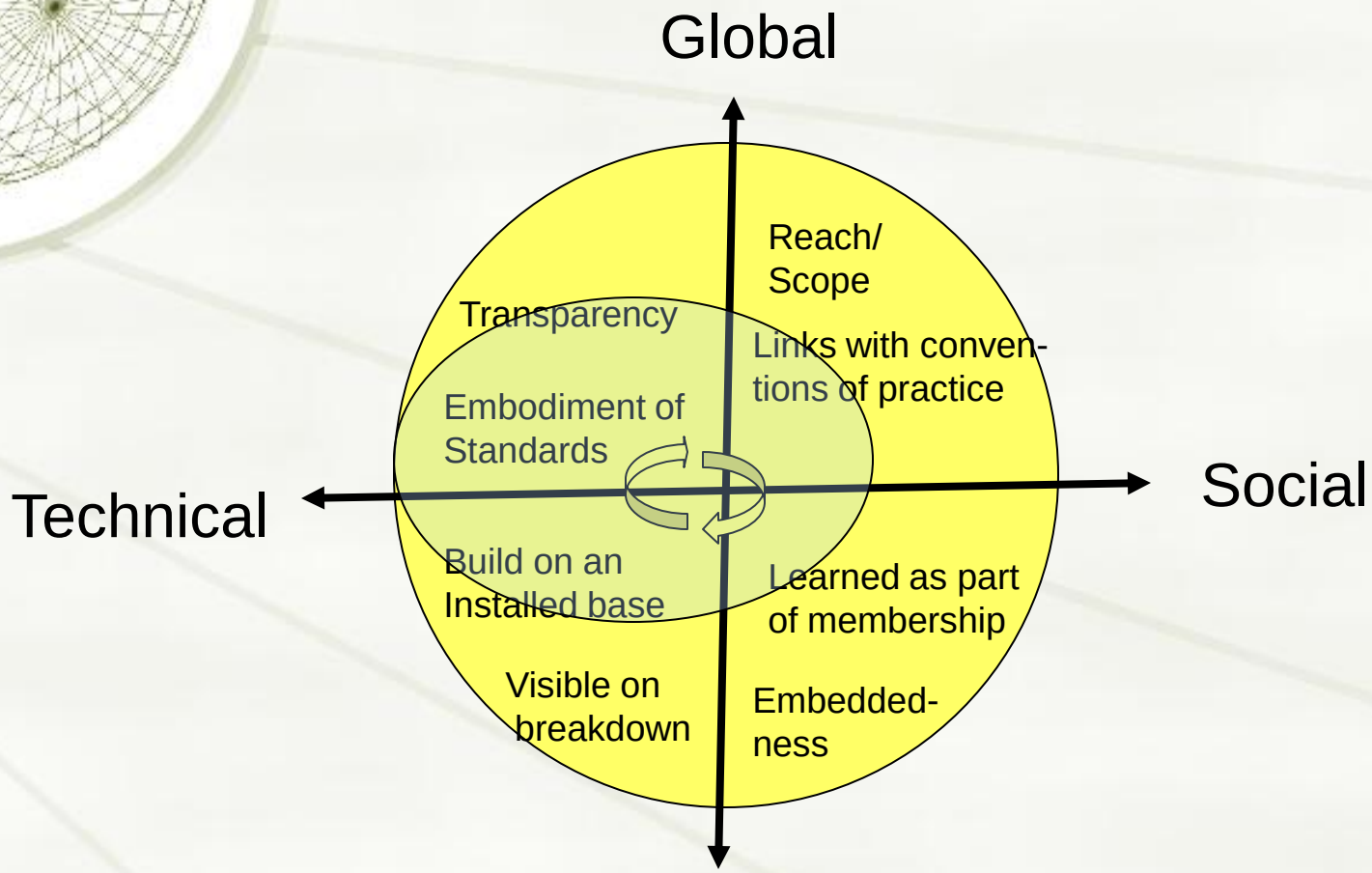
Barriers

- Can’t we free ride on cyber-infrastructure for e-science?
- Won’t Google do this for us?
- Won’t the Semantic Web fix all of these problems?
- How do we move from isolated projects to infrastructure?

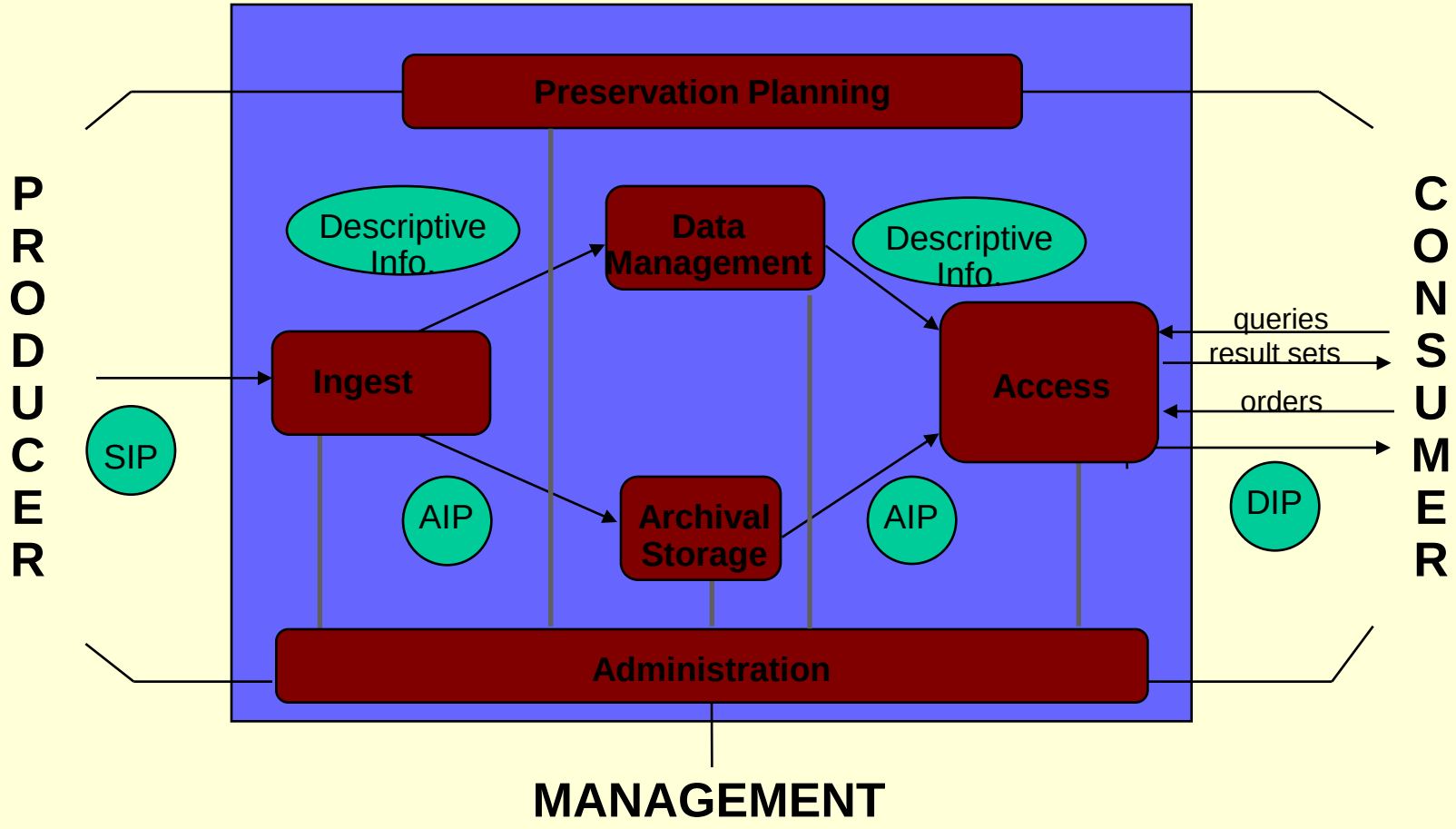
Identifying Gaps

- Most digital preservation research and development is centered on repositories
  - Architecture
  - Metadata
  - Tools
- Developments focus on the technical axis
- Many digital preservation efforts focus on activities within repositories
- Outreach to producers is limited to a subset of producer communities

Gaps in Infrastucture

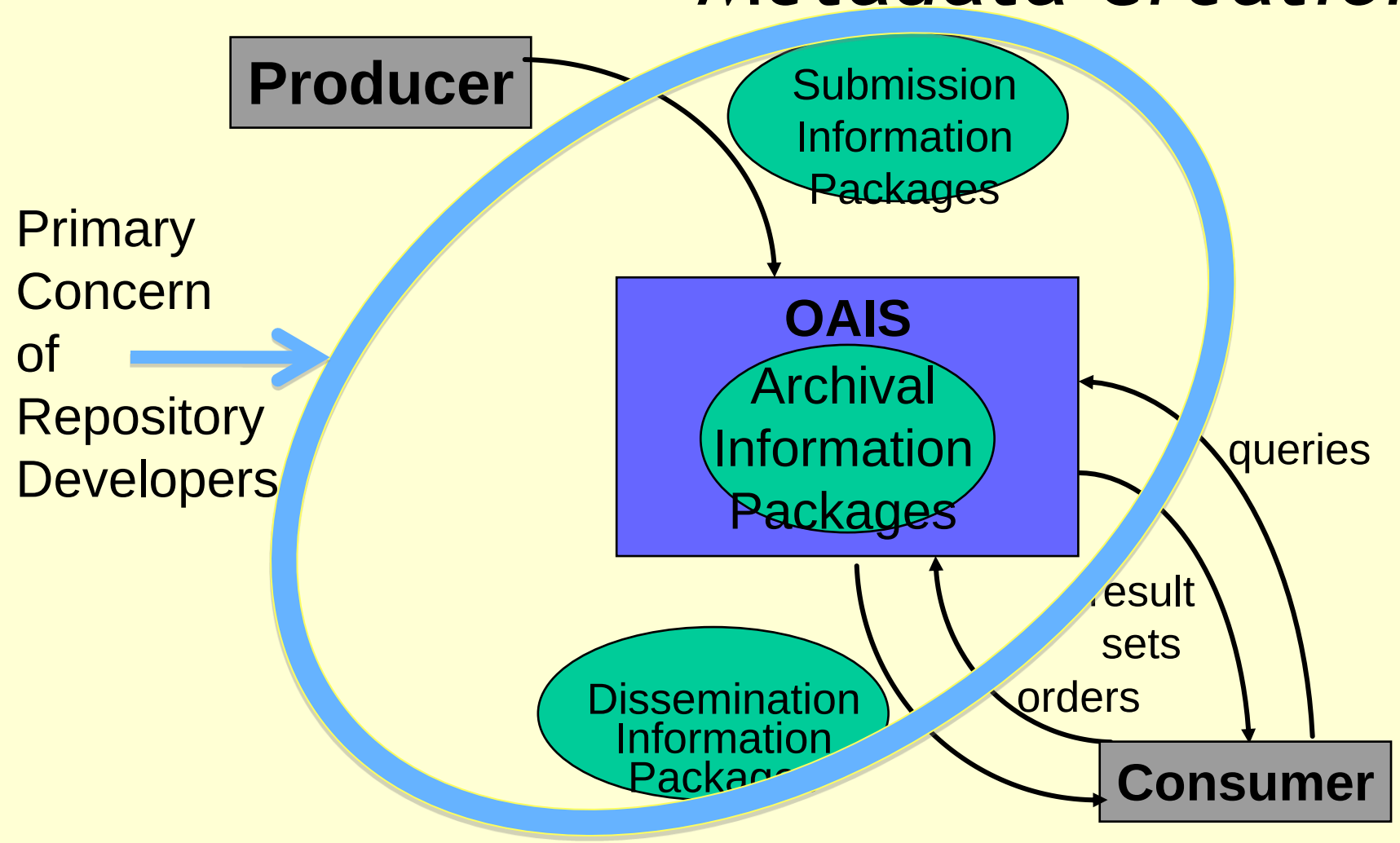


Scope of OAIS Activities

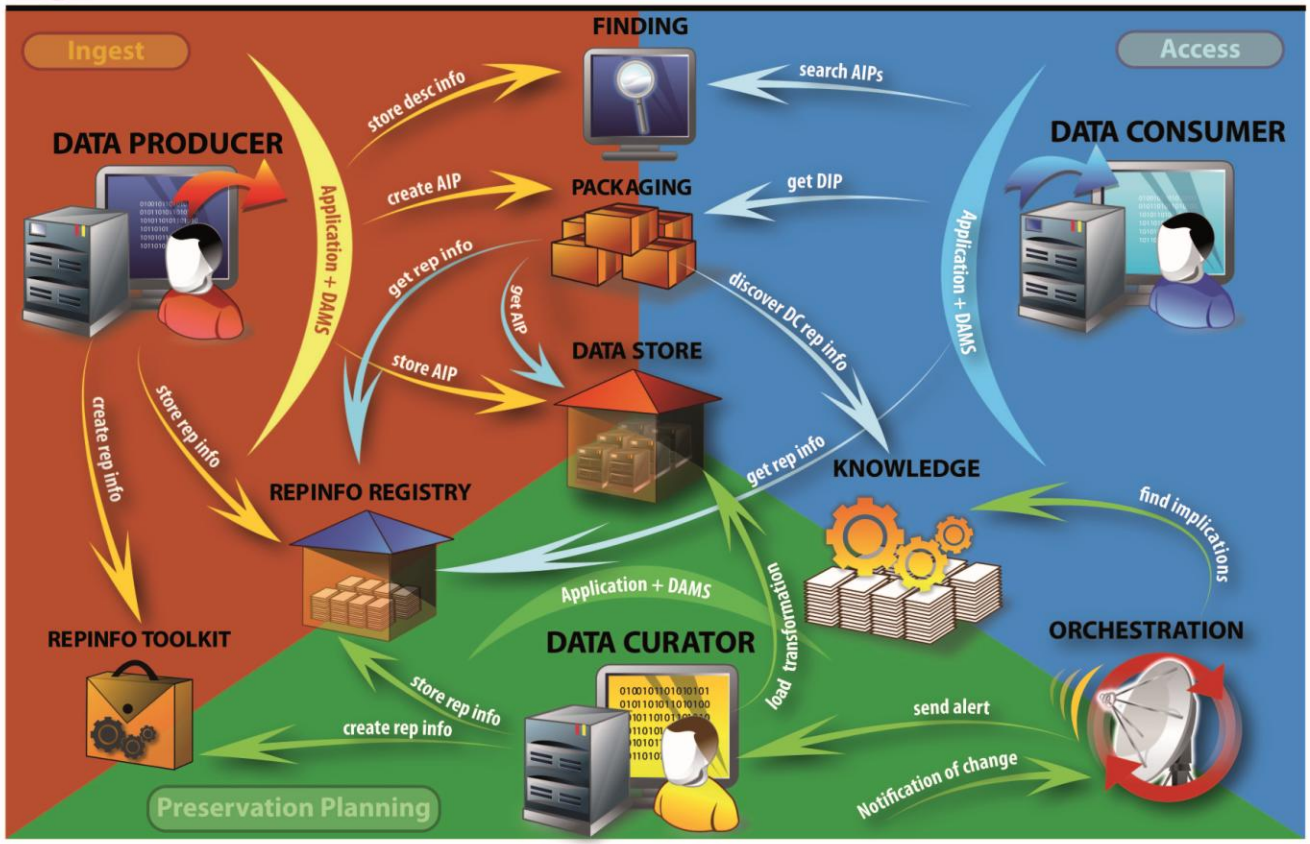


SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

Repository-Centered View of Metadata Creation



Repository-Centric View of Workflow



Identifying Gaps

- Interoperability between tools, standards and practices in producer communities and repository standards, tools and practices
- Two different workflows
  - Data production
  - Digital preservation

Identifying Gaps

- Social side of infrastructure
  - Reaching into more producer communities
  - Reaching more deeply into the data production process
- Provision for preservation becomes part of normal workflow
- Awareness and skill needed for preservation is learned as a part of collecting data, doing research, etc.

Bridging the Gaps

- How can we build infrastructure that unites the production of scientific data with long-term preservation?
- Technical Issues
  - Tools the interoperate between production and preservation environments
  - Workflows that begin in the production environment

Bridging the Gaps

- Social Issues
  - Can we embed preservation awareness in the scientific production environment?
  - Can we teach/learn good data practices as part of learning good research practice?
  - Can we extend models of good practice from one lab to the next? One discipline to the next?

Conclusion

- Building digital preservation infrastructure will require:
  - A long view of the information life cycle beginning at the point of creation (or before)
  - Embedding digital preservation requirements into systems and tools for producing information
  - Close attention to the fit between conventions of practice and preservation requirements