UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

5 November 2013
Lecture 20 – Midterm Review

The Course In One Slide

- To organize is to create capabilities by intentionally imposing order and structure
- We organize things, we organize information, we organize information about things, and we organize information about information
- If we think abstractly about these activities, we can see commonalities that outweigh their differences; We select, organize, interact with, and maintain resources
- We organize resources as individuals, in informal association with other individuals, or as part of a more formal institutional or business context
- We must recognize the profound impact of new technologies and their co-evolution with the nature of the organizing we do and the kinds of interactions that this organizing enables, but can't ignore the "classical" concepts and knowledge

Organizing Principles [1]

- ORGANIZING PRINCIPLES use properties or DESCRIPTIONS that are associated with the resources; organizing and describing resources are inherently interconnected activities
- Almost any property of a resource might be used as a basis for an organizing principle, and multiple properties are often used simultaneously
- The principles can also use collection-level properties

Organizing Principles [3]

- Other typical arrangements are based on ownership, origin, taxonomic, or “taskonomic” properties (usage frequency, correlated usage)
- Any resource with a orderable name or identifier can have alphabetic or numeric ordering
- Any resource with an associated date (creation, acquisition) can have chronological ordering
- Principles should be expressed logically in a way that doesn’t assume an implementation

Why We Describe Resources

- We describe resources so we can refer to them, organize them, and interact with them
- Each purpose might require different descriptions and different methods of using them
- Different resource domains can have characteristic or standard resource descriptions (or description categories)

Categories are "Equivalence Classes"

- Categories are sets or groups of resources or abstract entities that are treated the same
- This almost never means that every instance of the category is identical
- It only means that for some purpose we treat them in the same way

Categories are Models

- Defining categories as equivalence classes in this way should remind you of this definition:

  Models are simplified descriptions of a subject that abstract from its complexity to emphasize some features or characteristics while intentionally de-emphasizing others

- Categories are COGNITIVE and LINGUISTIC MODELS for applying prior knowledge

Principles for Creating Categories

- Enumeration
- Single Properties
- Multiple Properties
- Family Resemblance
- Similarity
- Theory-Based
- Goal-Derived

Distinguishing Categorization and Classification (1)

- Categories are EQUIVALENCE CLASSES - sets of resources, processes, and events that we treat the same
- A Classification (noun) is a SYSTEM OF CATEGORIES, ordered according to a PRE-DETERMINED SET OF PRINCIPLES and used to organize a collection of resources
- Classification (verb) is the process of systematically assigning resources to intentional (often institutional) categories in a classification system

Classification Schemes

- A HIERARCHICAL or TAXONOMIC scheme emerges when multiple resource properties are used by organizing principles; each property creates another level
- A scheme can be both HIERARCHICAL and ENUMERATIVE at the lowest level where resources are categorized
- A FACETED classification scheme uses multiple resource properties, but does not require every resource to have a value for every property and allows the properties to be considered in any order

Principles Embodied in the Classification Scheme

- Warrant: What is the justification for the choice of categories and their names?
  - Literary Warrant: Classify only the resources we have?
  - Scientific Warrant: Use expert categories and names
  - Use Warrant: Use categories and names from "ordinary" people
- Breadth and depth of classification hierarchy
- Degree of enumerativeness

Classification and Standardization (1)

- Classifications and standards both impose order on resources
- They both distinguish, explicitly or implicitly, between standard / appropriate / effective and nonstandard / inappropriate / ineffective ways of creating organizing, and using resources
- But this does not imply that a standard is a good one or that the best one will win a "standards war"

Description is Challenging

• People use different words for the same things, and the same words for different things - what would a "good" description be like, and how can it be created?

• Describing and organizing always (explicitly or implicitly) takes place in some context

• The context shapes which resource properties are important and the organizing principles that use those properties, introducing bias

The Vocabulary Problem

- People use a large variety of words for the same thing or concept
- Most people - especially system designers - are surprised by this because they think their own word choices are “intuitive” or "natural“
- The extreme variability of word selection is an inescapable fact that has its roots in the nature of language and categorization

More than a Controlled Vocabulary

- A controlled vocabulary is a standardized set of terms (such as subject headings, names, classifications, etc.) assigned by organizers / cataloguers / indexers of resources
- A metadata schema like the Dublin Core controls the kinds of assertions about resources that you can make in the first place
- Controlled vocabularies can be very useful requirements or recommendations about the values that are contained in the assertions (the information content of the assertion)

Implications for Vocabulary Design

- Choosing vocabulary terms, and precisely defining their semantics, is essential but impossible to do perfectly
- Your vocabulary must express what YOU intend, so you "look inward" -- analyze how you think about a domain
- You want others to understand what you mean, so you need to "look outward" -- analyze the terms used by your users, competitors, or subject matter experts
- You should reuse other vocabularies or thesauri if they exist, especially for any "horizontal" components, to improve transformability and interoperability
- But these three approaches may suggest different terms

Complications

- The properties of resources that are easiest to describe are not always the most useful ones, especially for information resources
- For non-text information resources this problem is magnified because the content is often in a semantically opaque format that cannot usefully be analyzed by people.
- Business strategy and economics strongly influence the extent of resource description
- But as bibliographic collections grow larger, we need more descriptions to satify the “frbr”

The Vision of the Semantic Web (1)

- In a classic 2001 paper Sir Tim Berners-Lee says:
- The Web can reach its full potential only if … data can be shared and processed by automated tools as well as by people...
- The Semantic Web will bring structure to the meaningful content of Web pages...
- For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently.

There Are No Modeling Shortcuts

- You might think that "modeling" means "writing a schema given a set of instances" or "inferring a schema from a single instance" (like you can with the "autogenerate" function in many XML editors)
- But schemas developed without a stage of conceptual design (other than very simple ones) are rarely very useful because they are too closely tied to the particular instances used, which may not be representative
- Sometimes schemas went through a stage of conceptual design but once the schemas are implemented the conceptual information isn't available to allow users to evaluate suitability

To Summarize…

- The original vision of the Semantic Web emphasized the creation of ontologies that robustly described the semantics of particular domains or contexts
- Lots of research was spawned by this vision, but the high bar of formal semantics and automated agents undoubtedly deterred “regular” people and firms from adopting it
- Semantic authoring can’t take off without tools that are simple to use as tools for designing and creating HTML pages

What's Different About Describing Multimedia?

- Sensory Gap
- Semantic Gap
- Proliferation Problem

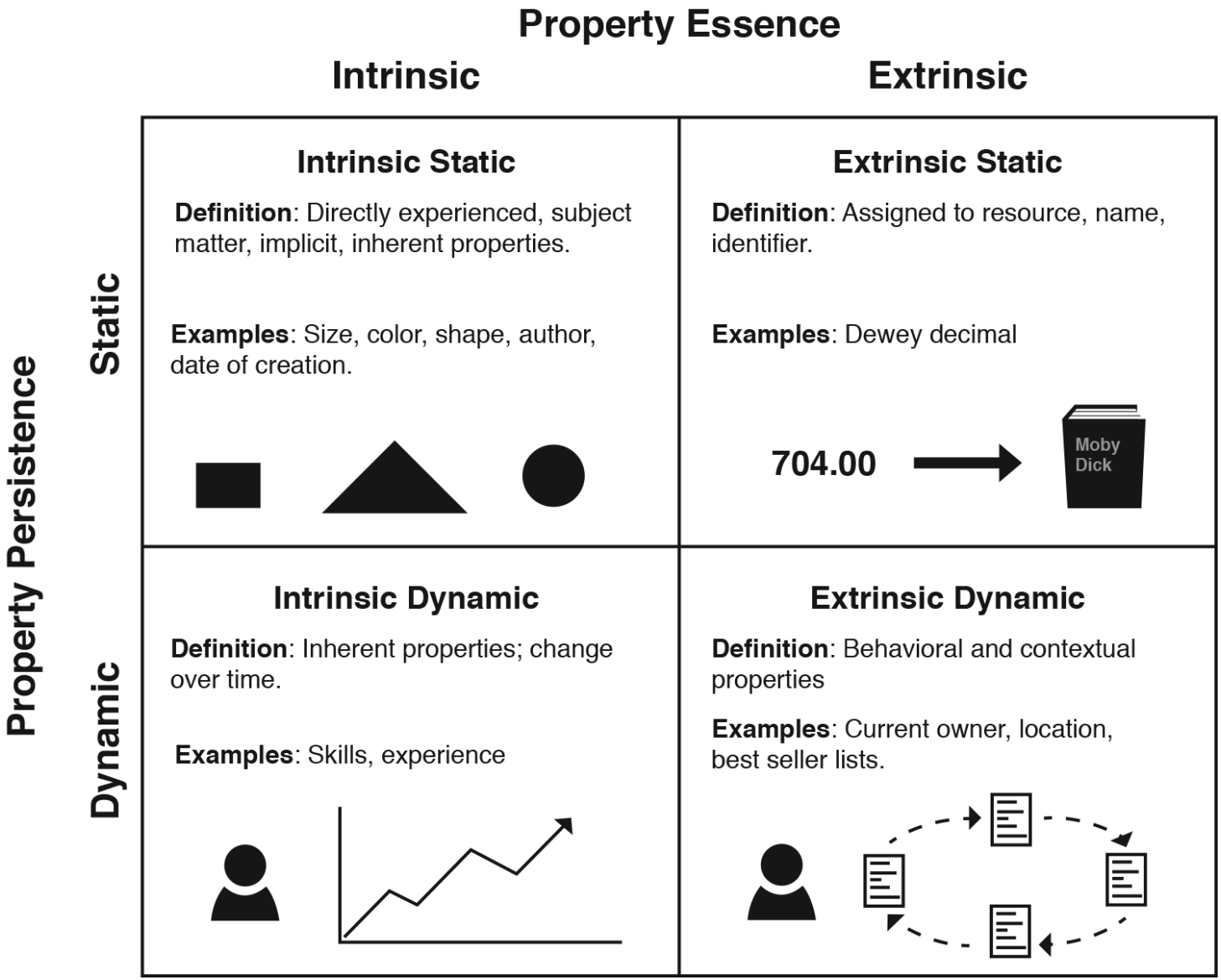
Are these “Gaps” New Problems?

- Museums face some of the same or similar problems in describing art works and artifacts:
- There may be many artifacts that represent the same "work" - this is like the "sensory" gap
- The materials or medium in which the artifact is embodied don't convey semantics "on their surface" - this is the semantic gap
- There may be so many artifacts of a particular type that some get only limited descriptions - this is like the proliferation problem

Some Problems May Be New

- The temporal structure of multimedia, especially video, mandates new descriptive vocabulary and new ways to identify meaningful components
- Video and music meet emotional/psychological needs that are more complex than those for "documents" - so the descriptions of the latter have to be able to address these needs
- People don't usually access or retrieve music or video "to satisfy information requirements“

Classifying Resource Properties



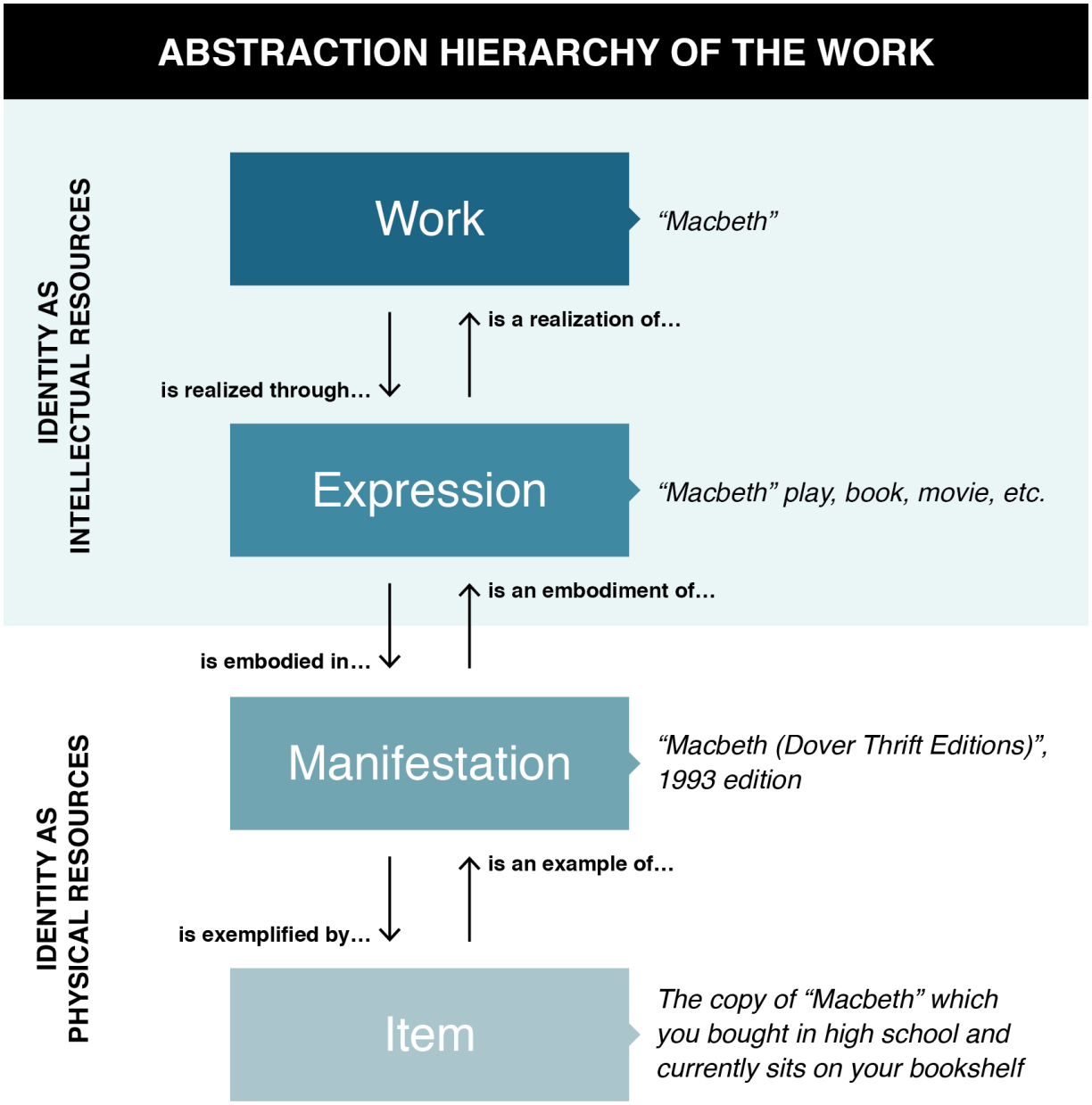
"Thing" vs "Type of Thing"

- Oops... we have been blurring the distinction between individual things or instances of things and classes of things
- We often say that that two objects are the "same thing" when we mean they are the same "type of thing"
- Identifying a resource as an instance is not the same as identifying the category or "equivalence class" to which it belongs

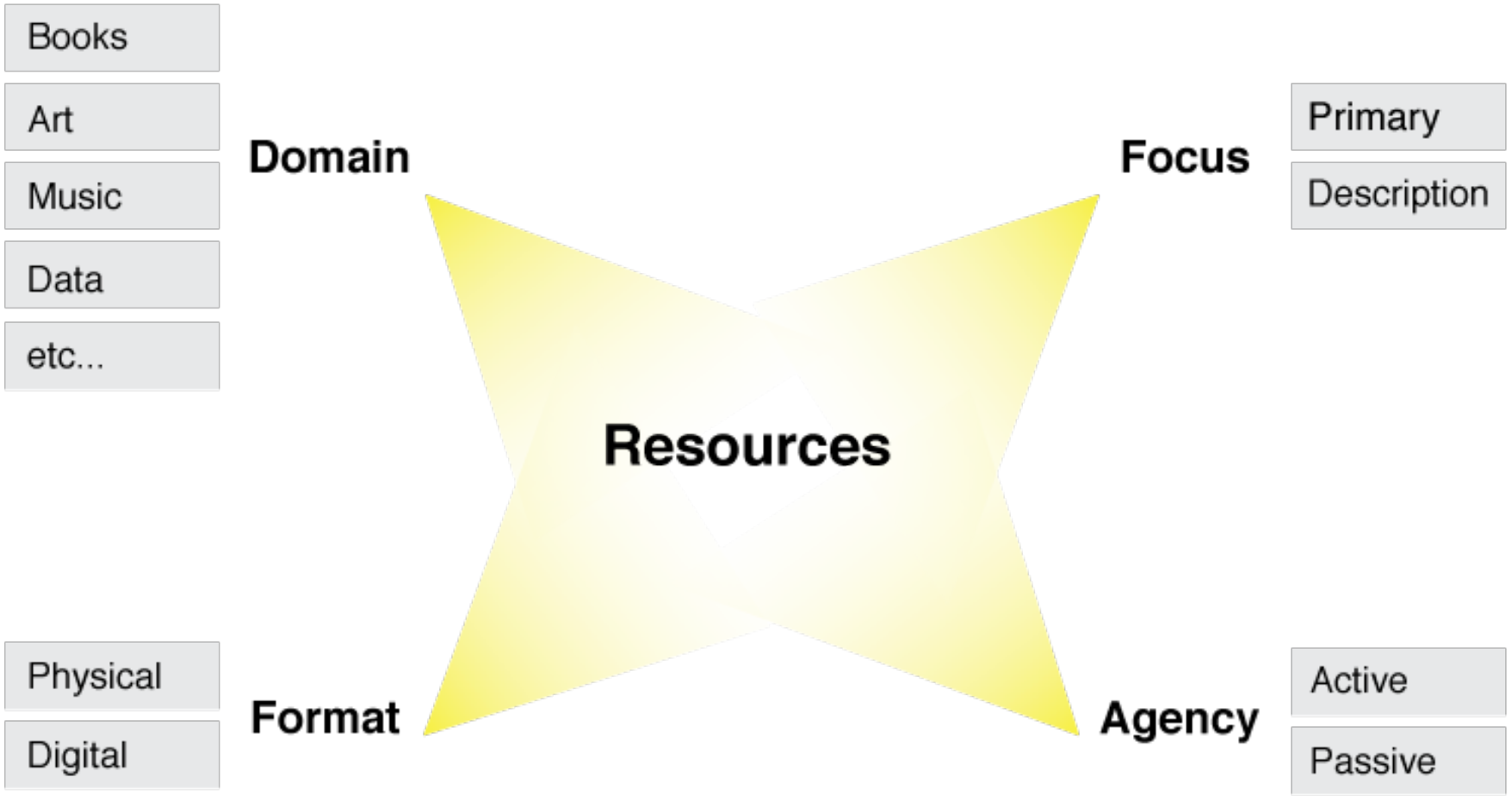
Two Aspects of “Thingness”

- Two separate aspects cut across the "thingness" distinctions:
  - Granularity
  - Abstraction

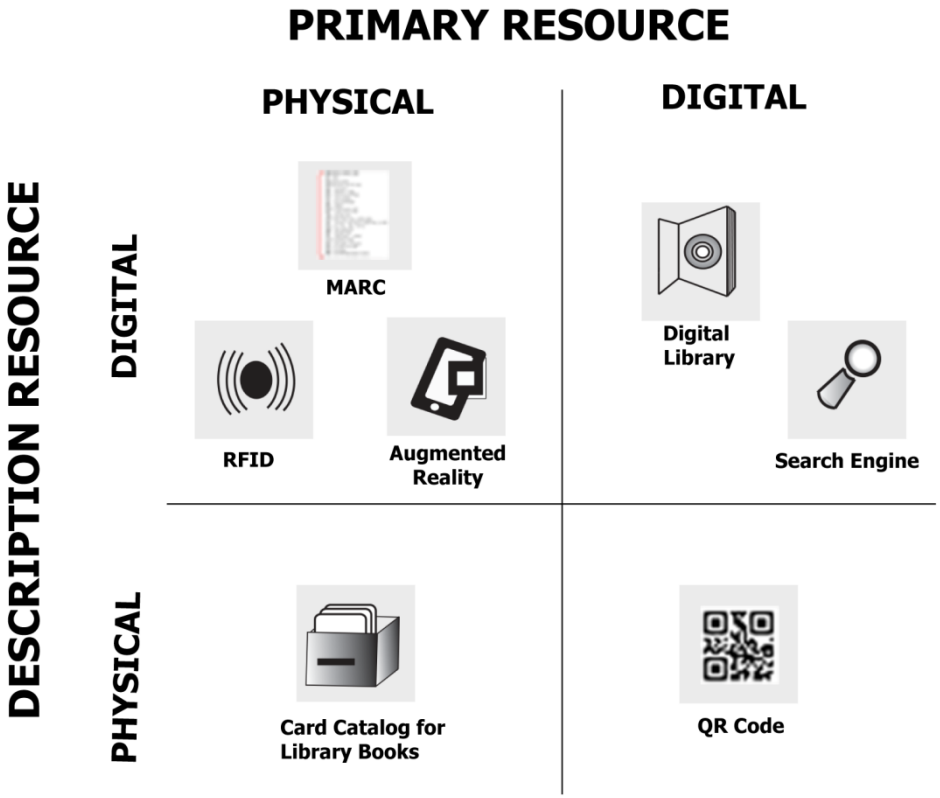
Instances vs. Types



Design Choices & Patterns for Resources



Format x Focus



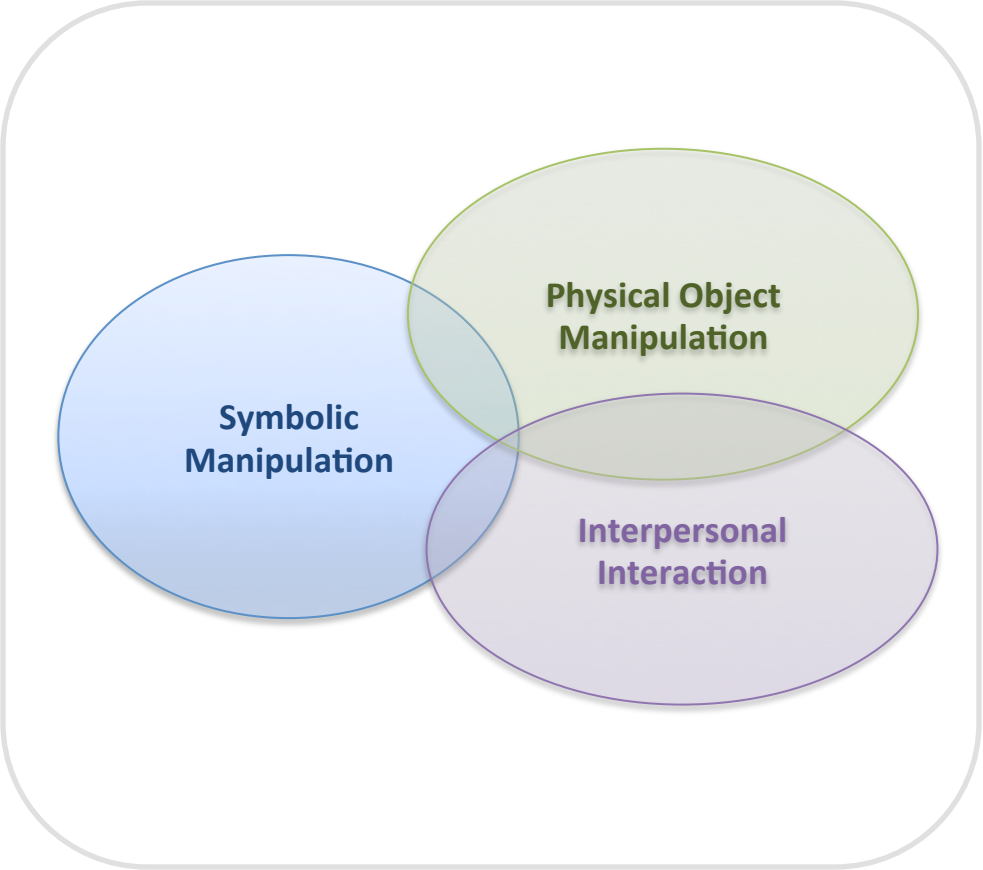
Interactions –The Why of Organizing Systems

- INTERACTIONS include any activity, function, or service supported by or enabled with respect to the resources in a collection or with respect the collection as a whole
- Interactions can include access, reuse, copying, transforming, translating, comparing, combining… anything that a person or process can do with the resources…

Interactions

- Some interactions can be enabled with any type of resource, while others are tied to resource types
- Interaction can be direct, mediated or indirect, or limited to interactions with resource copies or descriptions
- The supported interactions depend on the nature and extent of the resource descriptions and arrangement
- Different principles, or different implementations of the same organizing principles, determine the efficiency or effectiveness of the interactions

Interaction and Value Creation

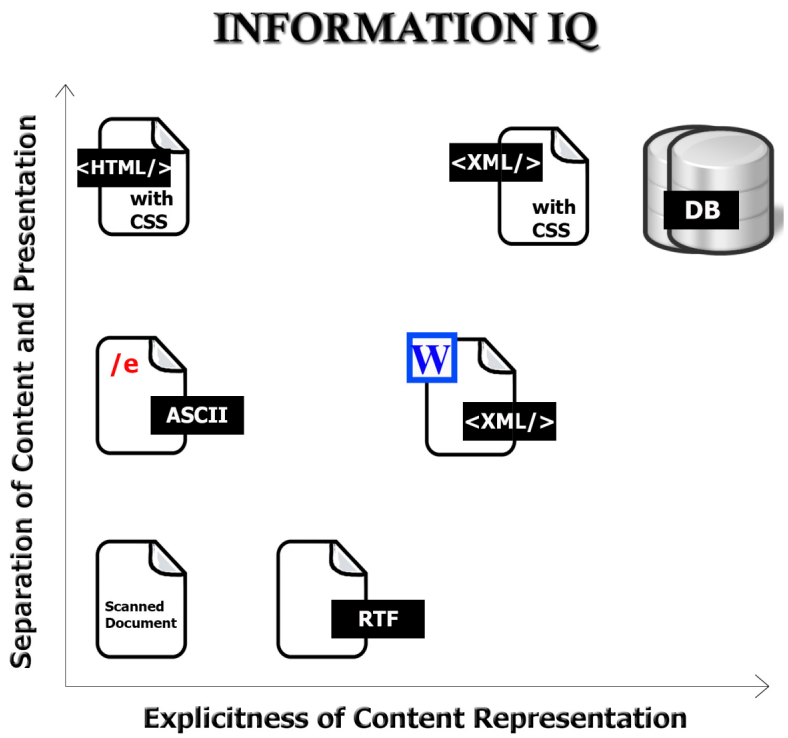


Interactions differ in the absolute and relative amounts of physical manipulation, interpersonal or empathetic contact, and symbolic manipulation or information exchange involved in the interaction

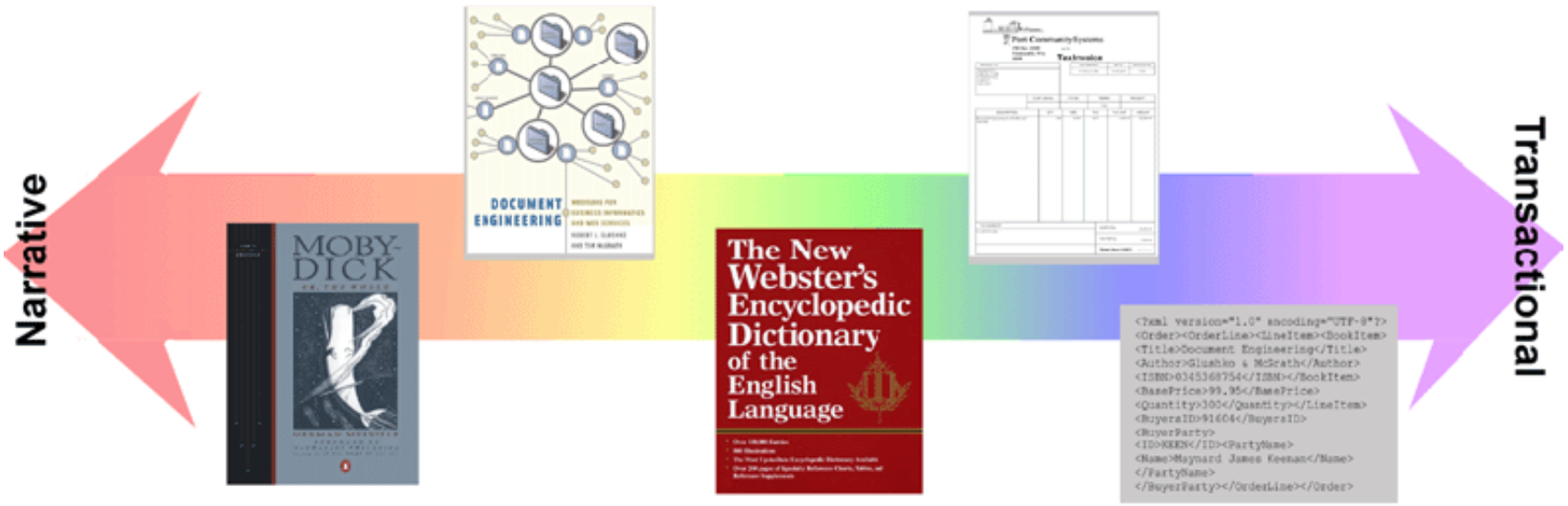
Apte, U. and Mason, R. Global Disaggregation of Information-Intensive Services. *Management Science* (1995).

Value Creation with “Smart” Resources

The variety and functions of interactions with digital resources are determined by the amount of structure and semantics represented in their digital encoding, in the descriptions associated with the resources, or by the intelligence of the computational processes applied to them

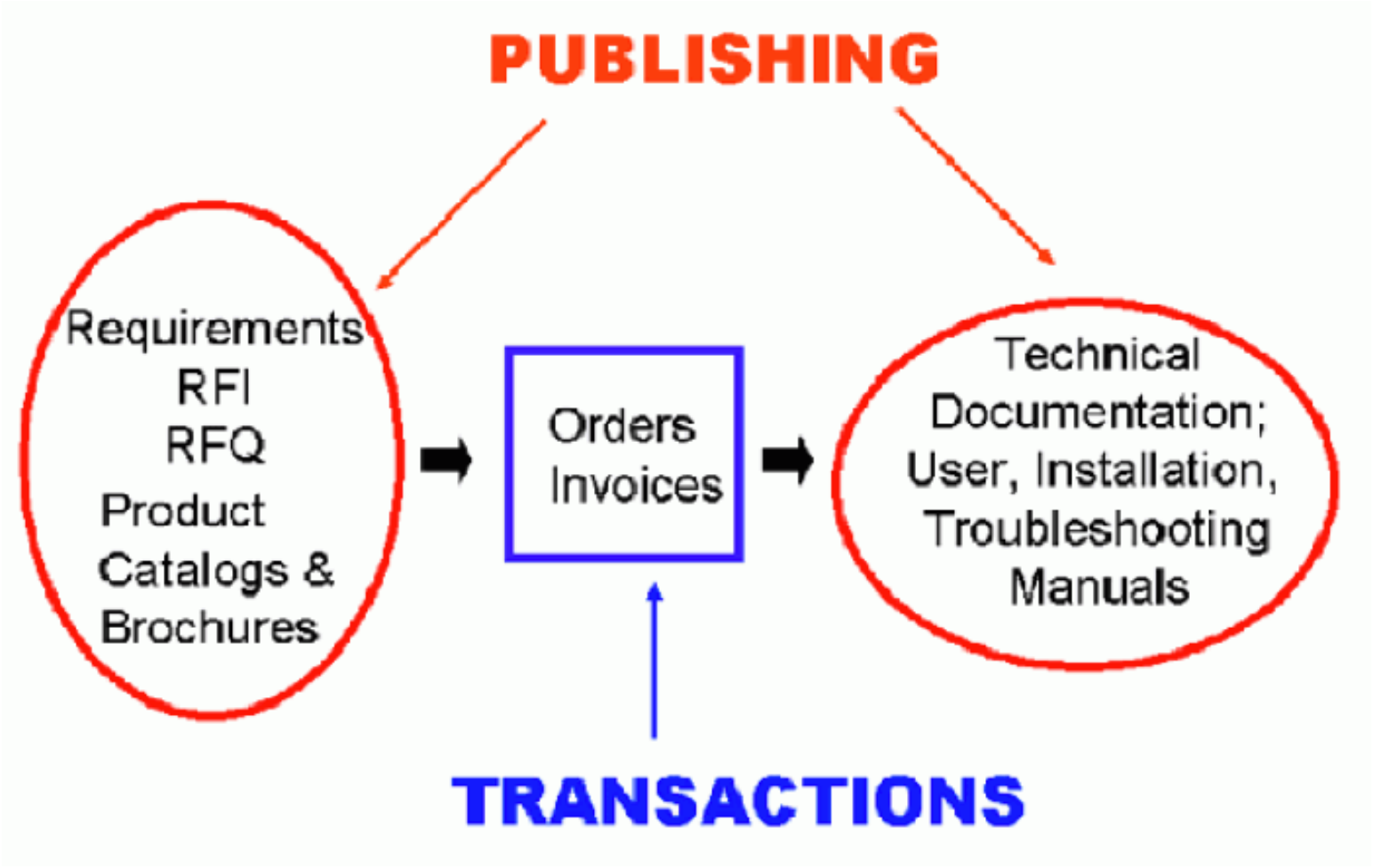


The Document Type Spectrum



From Glushko & McGrath, DOCUMENT ENGINEERING, MIT Press 2005

Mixing Data and Documents



Five Perspectives on Relationships (1)

- SEMANTIC: the meaning of the association
- LEXICAL: how the conceptual description of a relationship is expressed using words in a specific language
- STRUCTURAL: analyzes the patterns of association, arrangement, proximity, or connection between resources

Five Perspectives on Relationships (1)

- ARCHITECTURAL: emphasizes the number and abstraction level of the components of a relationship, which together characterize its complexity
- IMPLEMENTATION: how the relationship is implemented in a particular notation and syntax and the manner in which relationships are arranged and stored in some technology environment.

Defining "Relationship"

- “An association among several things, with that association having a particular significance”
- “Relationships are the stuff out of which information is made”
- The reason is an important part of the relationship
- Multiple relationships can exist among the same objects, so the order of the objects matters

The Structural Perspective on Relationships

- Analyzing the association, arrangement, proximity, or connection between resources without primary concern for their meaning or the origin of these relationships
- Sometimes structure is all we know...and sometimes we ignore what we know about relationship semantics to focus on the generic aspect of structural connectivity

Internal and External Structure

- Resources can have INTERNAL structure as well as EXTERNAL structure that connects them to other resources
- We often make arbitrary decisions about how the granularity with which we describe the internal structure of a resource
- The boundaries we impose to identify resources determines whether some structure is internal or external with respect to them

The Architectural Perspective {and,or,vs.} the Structural Perspective

- The architectural perspective is abstract and prescriptive
  - It defines what kinds of relationships can be created
- The structural perspective is concrete and descriptive one
  - It says "this is what exists" and describes the actual patterns of association, arrangements, proximity, or connection between resources"

Computing the Properties of Graphs

- Reachability – is there a path between any two nodes in the graph?
- Shortest path – if there are multiple paths between two nodes, which is the shortest?
- Centrality – which nodes are the most connected or have the average shortest paths to the other nodes?
- Subgraph discovery – are there sub-graphs that are completely contained in a larger graph?

The Conceptual Model of Social Tagging