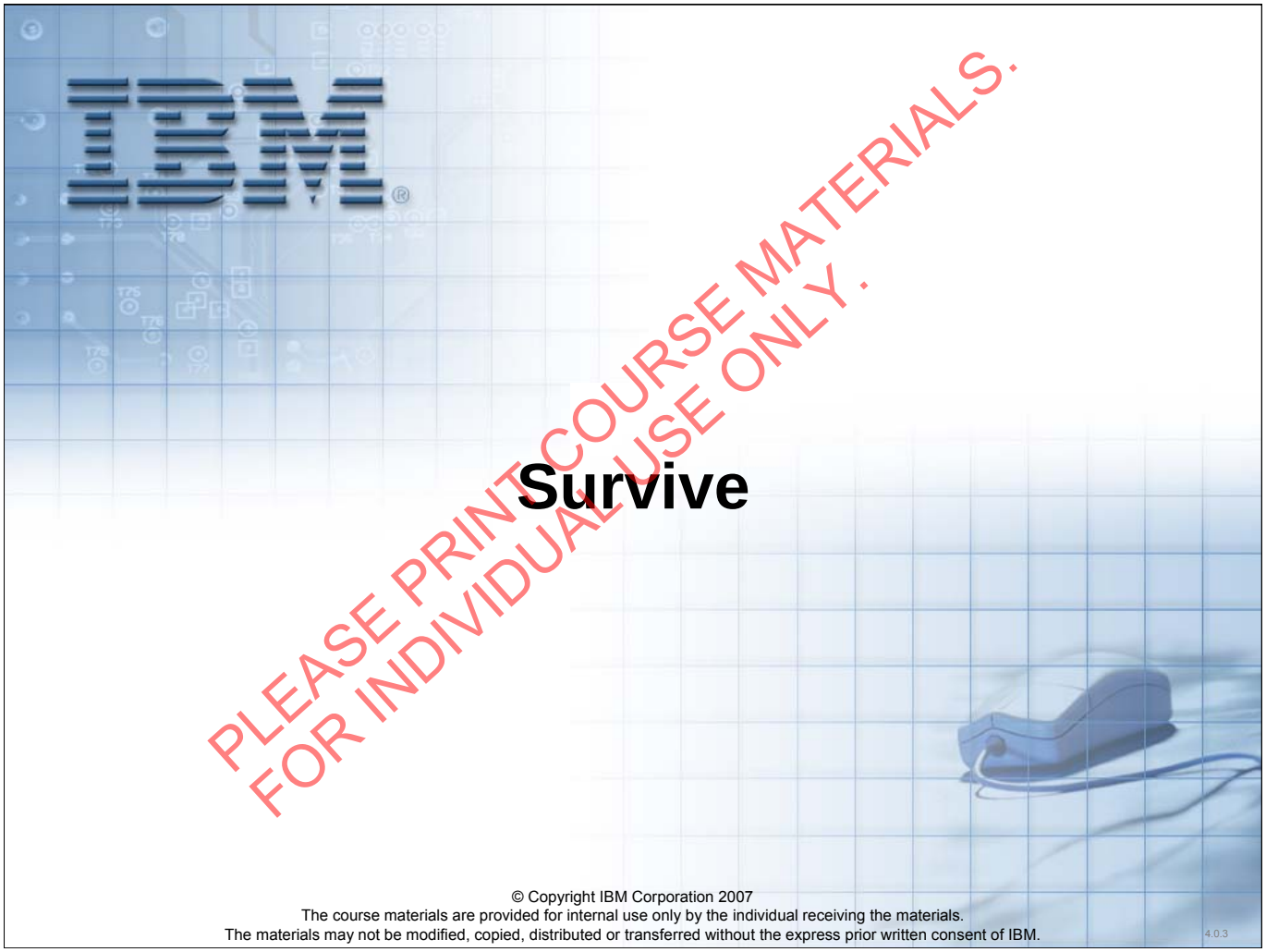


Unit objectives

- After completing this unit, you should be able to:
  - Identify Survive techniques
  - Describe implementation options
  - Define Survive rules
  - Build Survive job

Survive stage

- Point-and-click creation of business rules to determine "surviving" data – user decides how to survive data
- Performed at record or field level – very flexible
- Creates a single, consolidated record containing the "best-of-breed" data
- Cross-populates best available data
- Creates a cross-reference key
- Provides consolidated view of the data

Survive example

Survive Input (Match Output)

Group	Legacy	First	Middle	Last	No.	Dir.	Str. Name	Type	UnitNo.
1	D150	Bob		Dixon	1500	SE	ROSS CLARK	CIR	
1	A1367	Robert		Dickson	1500		ROSS CLARK	CIR	
23	D689	William	A	Obrian	5901	SW	74TH	ST	STE 202
23	A436	Billy	Alex	O'Brian	5901	SW	74TH ST		
23	D352	William		Obrian	5901		74 ST		#202

Survived Consolidated Output

Group	Legacy	First	Middle	Last	No.	Dir.	Str. Name	Type	Unit	No.
1	D150	Robert		Dickson	1500	SE	ROSS CLARK	CIR		
23	D689	William	Alex	O'Brian	5901	SW	74TH	ST	STE	202

Cross-Reference File

Group	Legacy
1	D150
1	A1367
23	D689
23	A436
23	D352

Survive rules

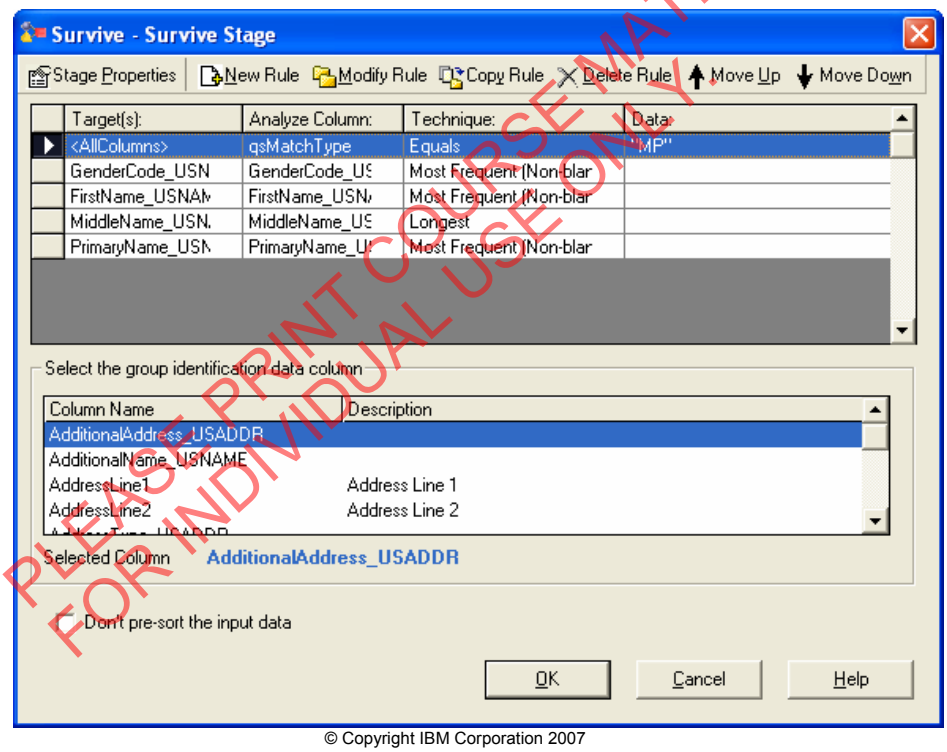
- A rule contains a condition and a set of target fields
  - When the condition is met the field becomes a candidate for the “best”
  - All records in a group are tested against the condition
  - The “best” populates the target fields
- Multiple targets are permitted for the same rule

Survive rules

- Custom Rule
  - Build your own logical expression
  - Comparison (=, !=, <, > ,<=, >=)
  - Logical (and, or, not)
  - Indicate the current and best records with the following notation
    - c.field indicates the current
    - b.field indicates the best
  - Parentheses ( ) can be used for grouping complex conditions
  - String literals are enclosed in double quotation marks, such as “MARS”.
  - A semicolon (;) terminates a rule.

Building survive rules

- Survive Rules Definition screen lets you easily build, delete and manage survivor rules



Survive techniques

- Pre-defined Techniques
  - Source
  - Recency
  - Frequency
  - Most complete (longest string)
- User-specified logic

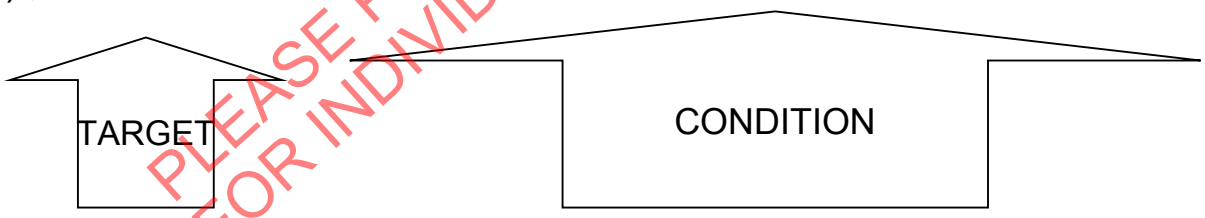
Target fields

- Fields you want to write to the output file
- Populated based on meeting the conditions of the survivor rule(s)
- Fields not listed as targets are excluded from the output file
- May have multiple targets for each rule

Example: complex survive rule

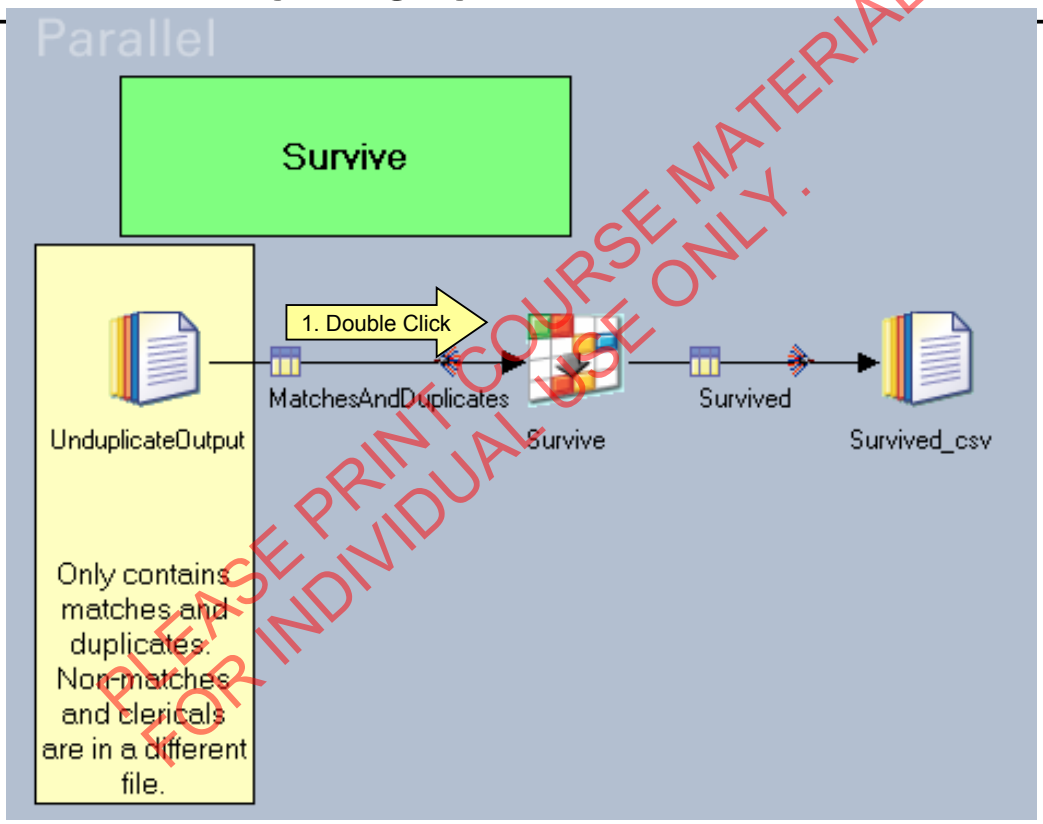
- The following rule states that FIELD3 of the current record should be retained if the field contains five or more characters and FIELD1 has any contents.
- The prefix of b. indicates the current “best” record
- The prefix c. indicates the current record testing against the survivor rule

FIELD3: (SIZEOF (TRIM c.FIELD3) >= 5) AND (SIZEOF (TRIM c.FIELD1) > 0) ;

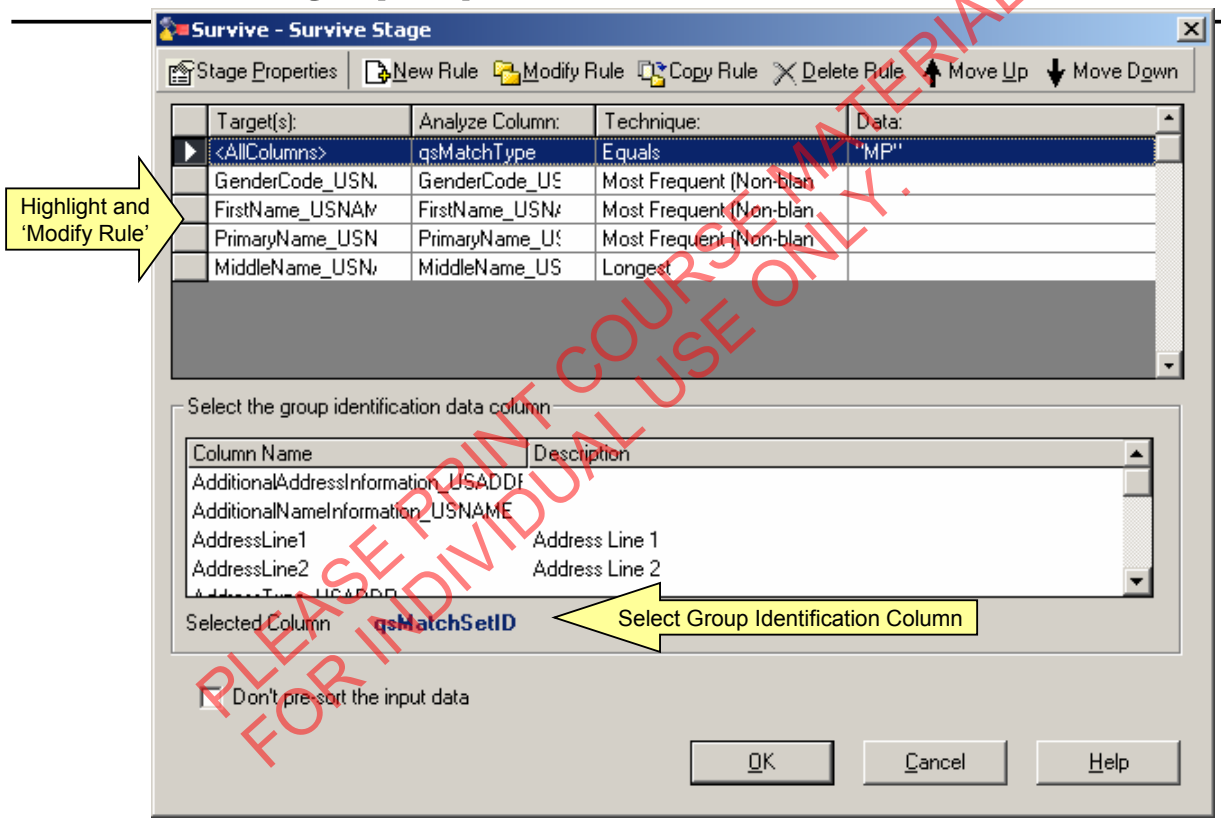


Survive Implementation

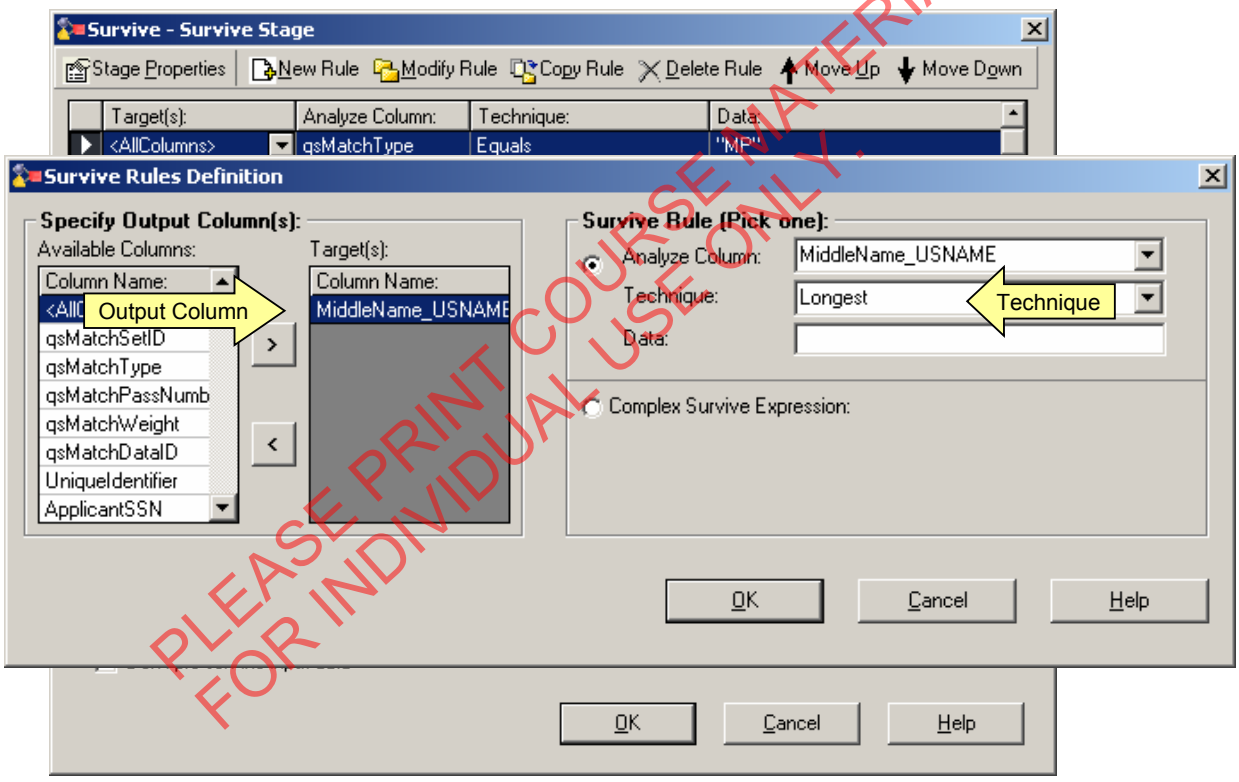
Survive QualityStage job



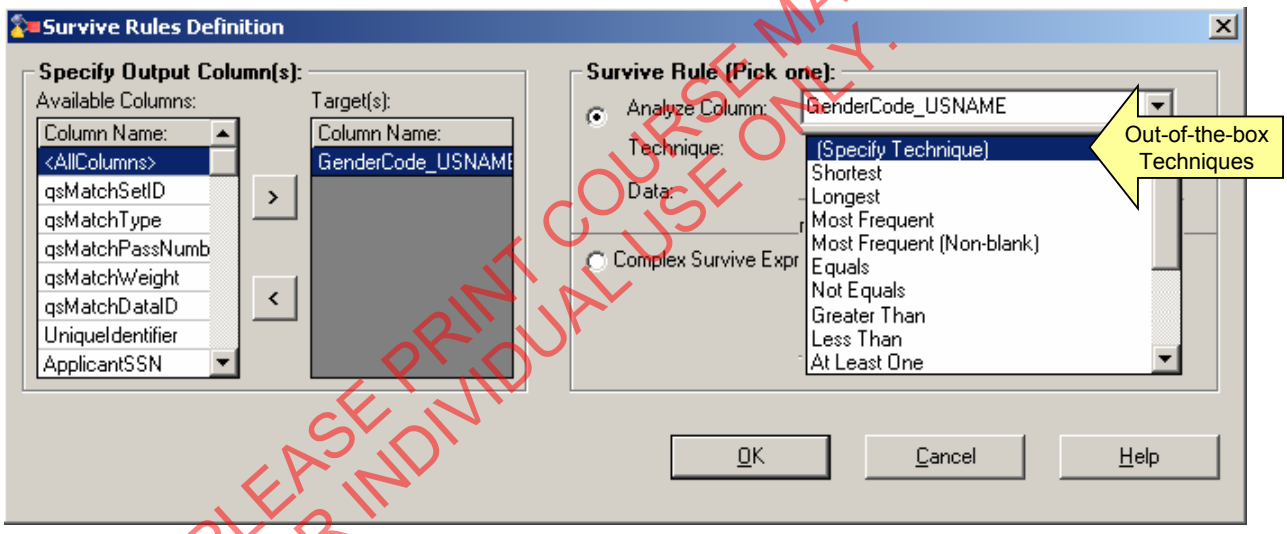
Survive stage properties



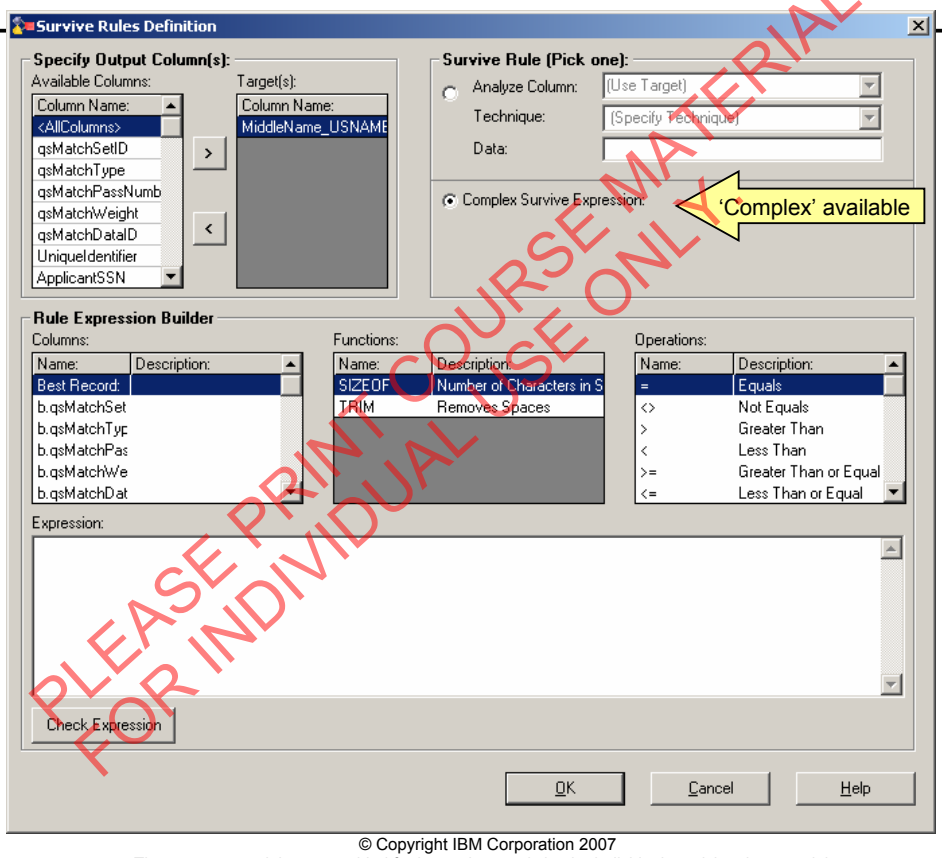
Survive stage properties



Survive stage properties



Survive stage properties



Checkpoint

1. (T/F) Survivorship can allow more than one record to survive.
2. (T/F) Survivorship rules deal with the complete record only.
3. Name three survive rules.

Checkpoint solutions

1. (T/F) Survivorship can allow more than one record to survive.
   Answer: False
2. (T/F) Survivorship rules deal with the complete record only.
   Answer: False
3. Name three survive rules.
   Answer: most recent record, longest non-blank, most frequent non-blank

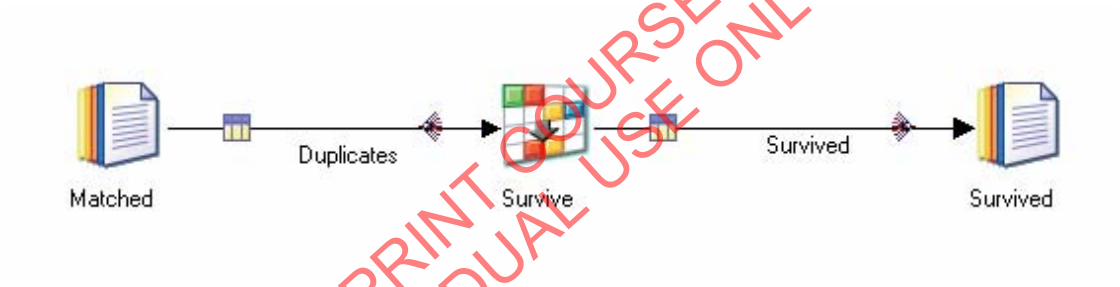
Unit summary

Having completed this unit, you should be able to:

- Identify Survive techniques
- Describe implementation options
- Define Survive rules
- Build Survive job

Exercise 21: Survivorship job

- Build survivorship job using the match specification



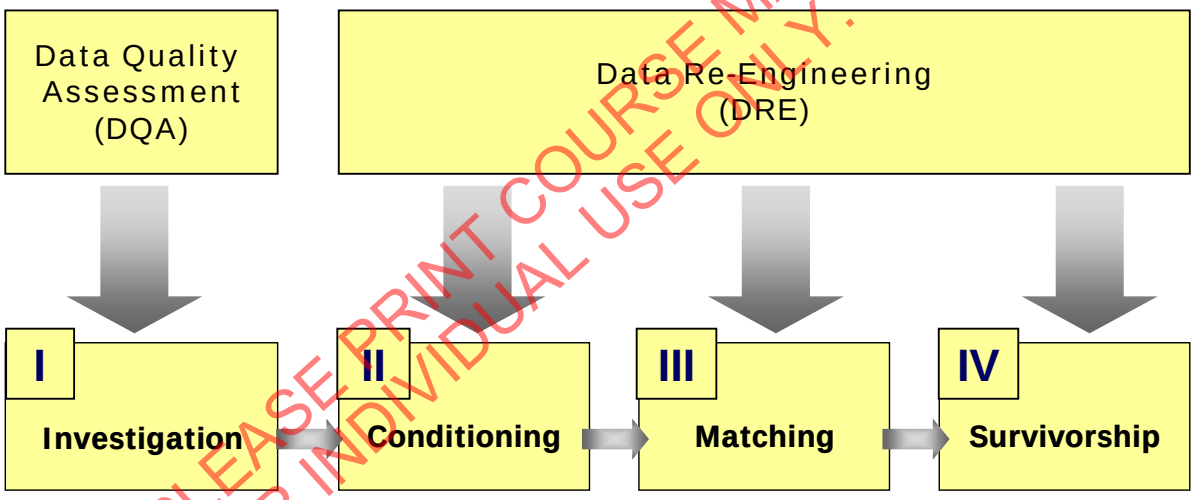

IBM
Special Topics
PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.
© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.
4.0.3

QualityStage Methodology

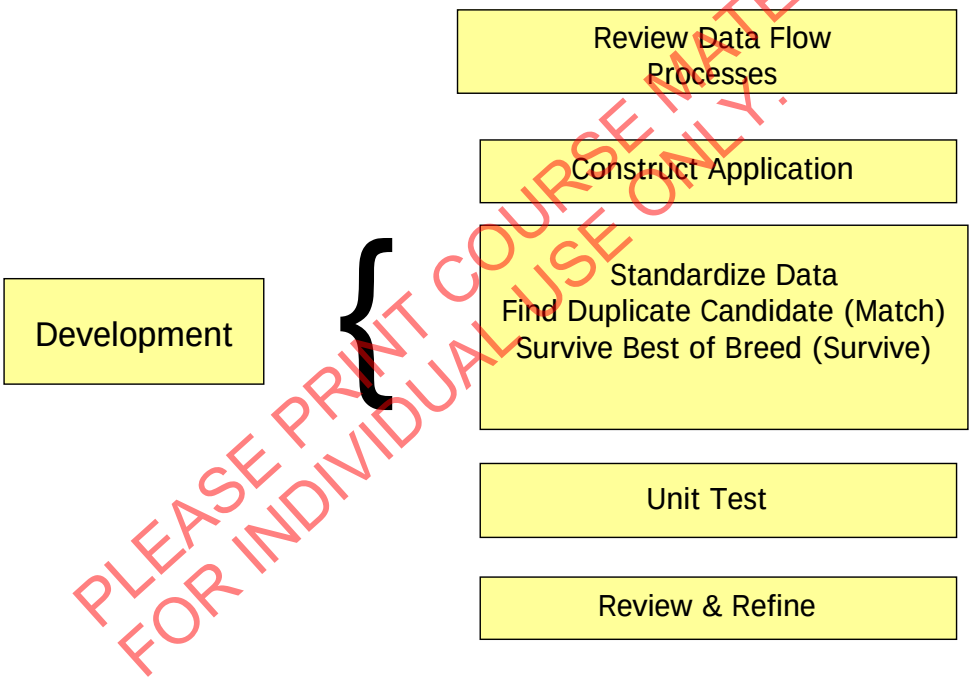
Unit objectives

- After completing this unit, you should be able to:
  - Describe the Data Quality Methodology
  - Identify which phase contains the Data Re-engineering Methodology steps.

Methods

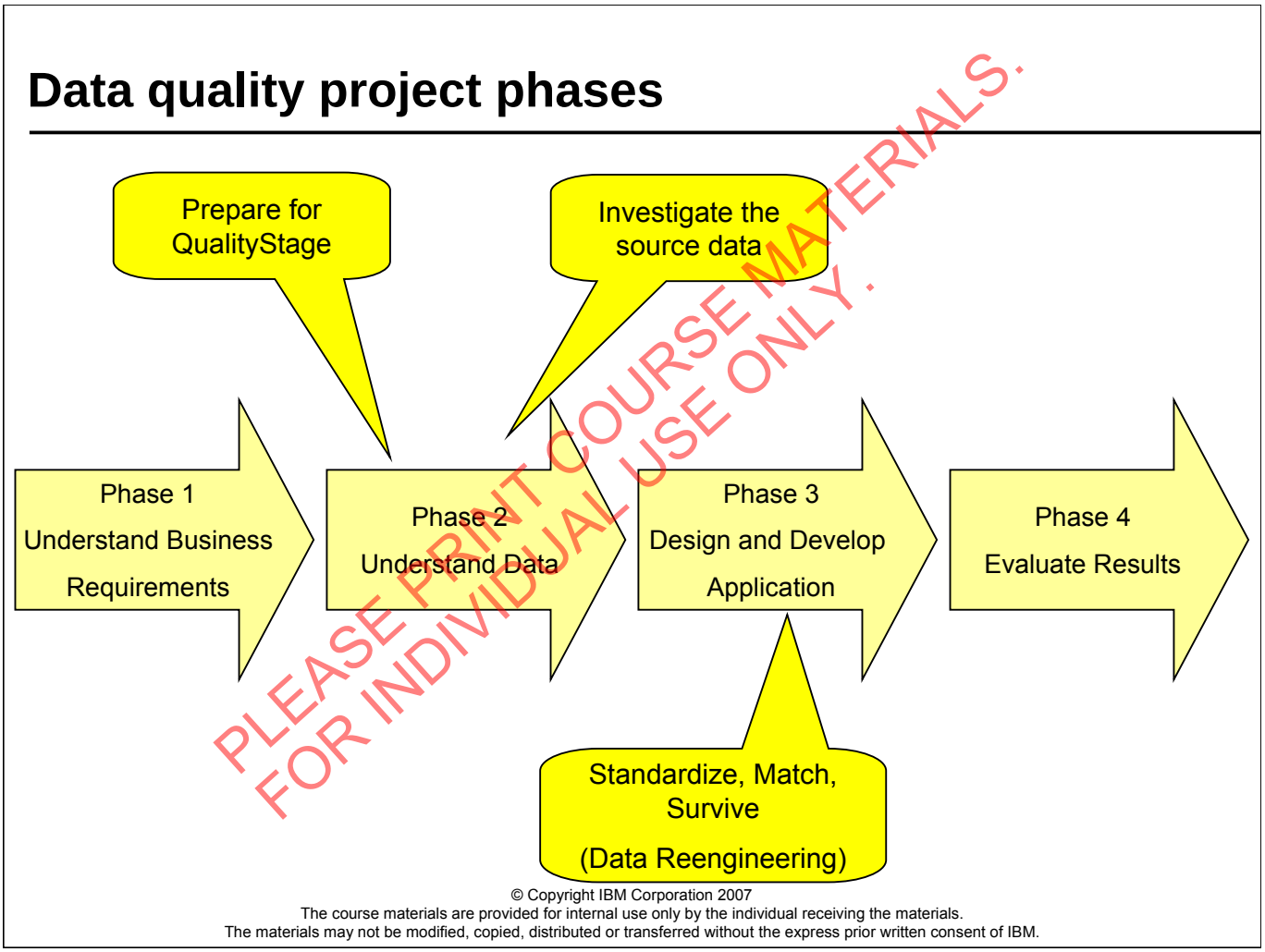


Project lifecycle: Development



Workflow for a QualityStage project

- Understand the business requirements
  - Business goals determine the data reengineering requirements
- Understand the data
  - Necessary to understand the size and complexity of the project before creating reengineered data
    - Investigate stage
- Design and develop the data reengineering application
  - Align data
    - Standardize stage
  - Link records
    - Match stage
  - Consolidate data
    - Survive stage
- Evaluate results
- These workflow activities are organized into a four phase methodology



Understand Business Requirements :

- Meet with the client, establish project requirements and document the results.

Preparation for QualityStage :

- Translate business requirements into data reengineering actions.
- Install QualityStage environment.

Investigate :

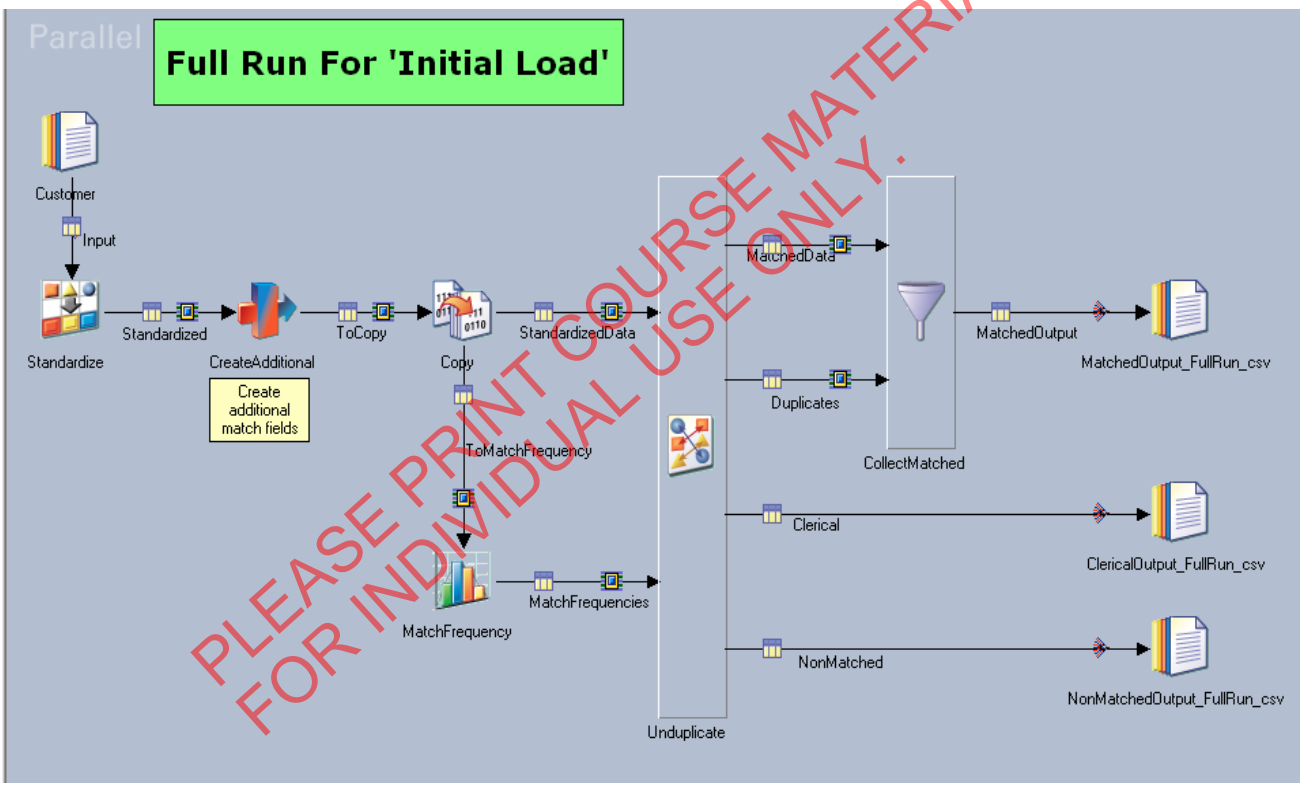
- Determine format and structure of source data.
- Understand the content of source data.

Data reengineering glossary

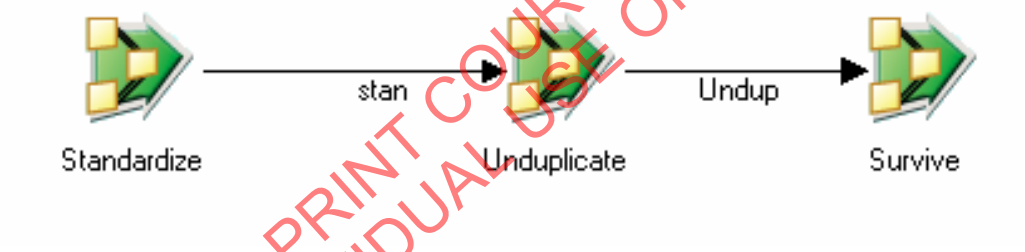
Business Term	*QualityStage*
Data Assessment	Investigate
Data alignment	Standardization
Data harmonization	Match
Consolidation	Survive

Full Run

Full run – single job



Full run using DataStage job sequencer



QualityStage Migration Tool

QualityStage migration tool objectives

- Overcome the 7.x integration issues - - no need to run via the QSDS plug-in
- Allows the inclusion of Legacy QualityStage jobs (or pieces) in a QS8 job without having to re-implement the job design
- Expands the connectivity options of QualityStage jobs by running in the QS8 environment
- Provides the opportunity to improve performance of QualityStage jobs by optionally converting to native QS8 stages for parallel execution (as opposed to running sequential, wrapped stages)

QualityStage Migration Tool – Overview

- The QualityStage Migration Tool (QSMT) provides the ability to migrate QualityStage 7.5 jobs and Standardization Rule Sets to the QS8 environment.
- QSMT analyzes the QS 7.5 server project directory to construct “dsx” files which can be imported into the QS8 common repository using the DS & QS8 Designer’s “import” facility

QualityStage migration tool – overview

- QSMT functionality offers three types of QS 7.5 **object** migrations:
  - QS 7.5 Standardization Rule Set
  - QS 7.5 job in combined mode
  - QS 7.5 job in expanded mode

- Two modes for migrating **jobs** to QS8:
  - Combined Mode
    - Use when you need to take a legacy process and just run it in QS8
    - Allows control before and after the legacy process
    - Will always run after importing without any manual tuning
  - Expanded Mode
    - Use when you need to add QS8 operators within a migrated process
    - May require some manual tuning to run

Rule set migration

- The QSMT has the ability to migrate Standardization Rule Sets in one of two ways:
  - Explicitly - - you may specify the rule set you want to migrate
  - By job dependency - - you may migrate all Rules associated with a particular job

Note: Regardless of the migration mode, all migrated rules will have the new naming convention of : *QS-7.5-Ruleset-Name_QS-7.5-Project-Name*

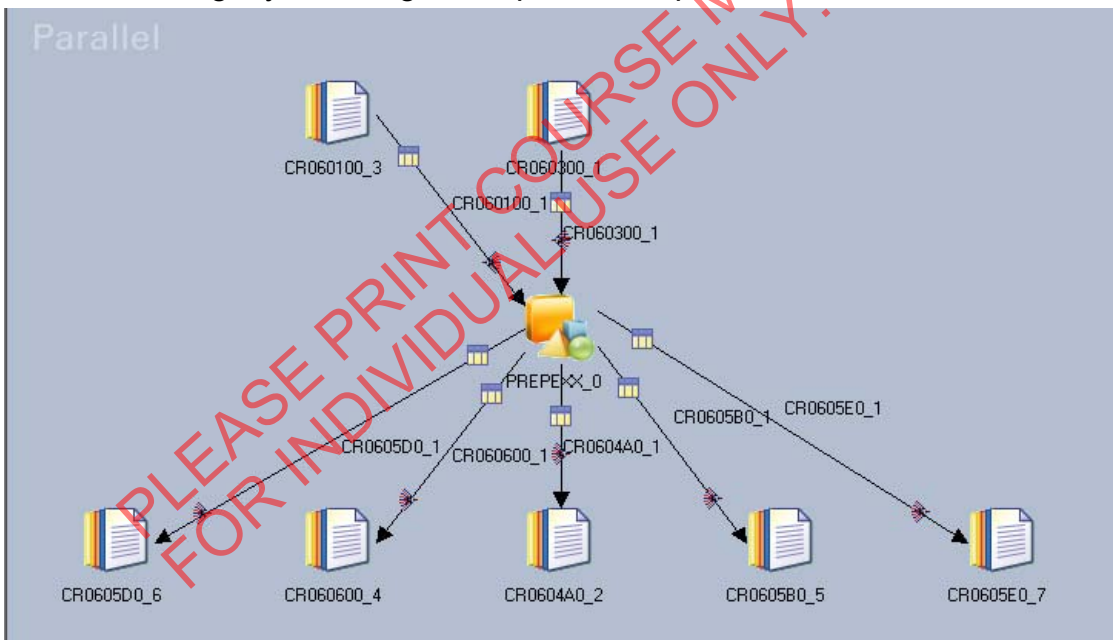
Combined mode migration

- Use this mode to get a legacy QS job up and running in QS8 with as little effort as possible. Jobs will import and run without modifications
- After importing, a migrated job will appear in the “Jobs” folder of the repository view in the QS/DS 8 Designer client
- Jobs are renamed by QSMT within the QS8 package to minimize name collision
- The new job name has the following naming convention:

  QS-7.5-Job-Name_QS-7.5-Project-Name

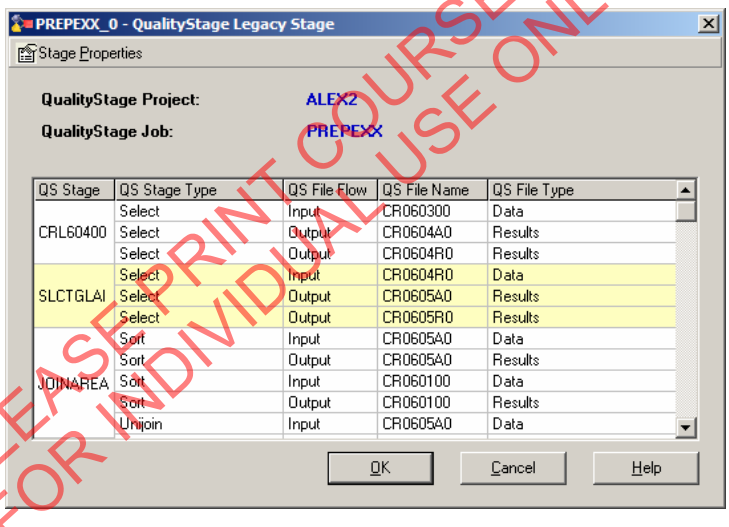
QSMT – combined mode migration

- The job consists of a single instance of the QS 8 Legacy Job stage, together with some number of DS Sequential File stages, which are linked to the Legacy Job stage as inputs or outputs



QSMT – combined mode migration

- All the QS stages run under the control of the single Legacy Job stage in Combined Mode
- The list of operations can be seen by opening the Legacy stage



- File IO to external files is performed by the Information Server Sequential File stages

QSMT – combined mode & running a QS8 job

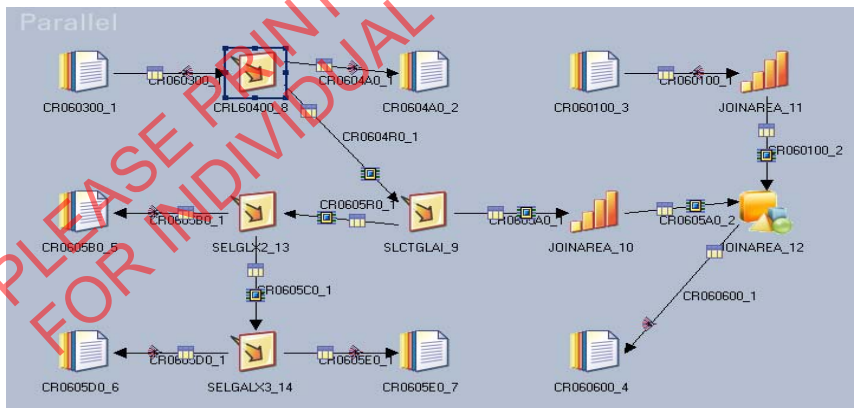
- Once imported, Legacy jobs are run the same as any other QS8 job
  - Prior to compiling, be sure any required rule sets are provisioned to the server
  - Run as you would any other QS8 job

QSMT – combined mode job results

- When running with a single node, output should exactly match that produced by QS 7.5 in stream mode
- When running with multiple nodes, output should exactly match that produced by QS 7.5 in PX mode when run with the same number of nodes

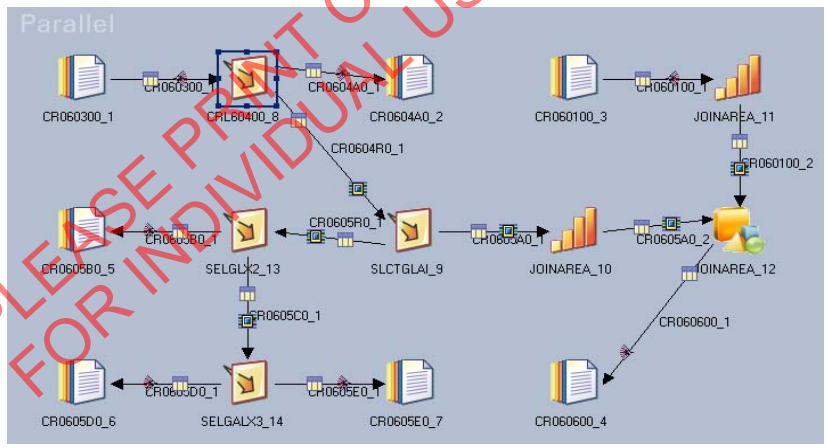
QSMT – expanded mode

- Use to re-implement the job in the QS8 environment
- After importing, a migrated job will appear in the “Jobs” folder in the same way as in Combined Mode
- The job consists of one or more stages for each 7.5 stage, plus DS PX Sequential File stages, linked to represent the 7.5 job flow. For complex jobs, stages may need to be reorganized to improve readability



QSMT – expanded mode

- There will always be at least one stage for each QS 7.5 stage
- The stage on the canvas may be a Legacy Job stage that runs a single QS 7.5 operation
- The remaining QS7.5 operations are converted to a variety of stages



QS stage migration reference table

QS 7.5 Stage Type	QS8 Stage Type	Conditions
Abbreviate	Legacy Job	Always
Build	Legacy Job	Always
Collapse	Legacy Job	Always
FFC	Copy	"Delimited text" used in 7.5 stage
FFC	ODBC Enterprise	"ODBC" used in 7.5 stage
Investigate	Legacy Job	Always
Match*	Legacy Job	Always
Multinational Standardize	MNS	Always
Parse	Legacy Job	Always
Select	Legacy Job	"Merge" used in 7.5 stage
Select	Filter	"Split", "Accept", or "Reject" used in 7.5

* Currently working on converting Match specifications for GA

QS stage migration reference table

QS 7.5 Stage Type	QS8 Stage Type	Conditions
Sort	Sort	Always
Standardize	Standardize	Always
Survive	Legacy Job	If target columns overlap
Survive	Survive	If target columns do not overlap
Transfer	Legacy Job	Always
Unijoin	Legacy Job	Always
WAVES	WAVES	Always

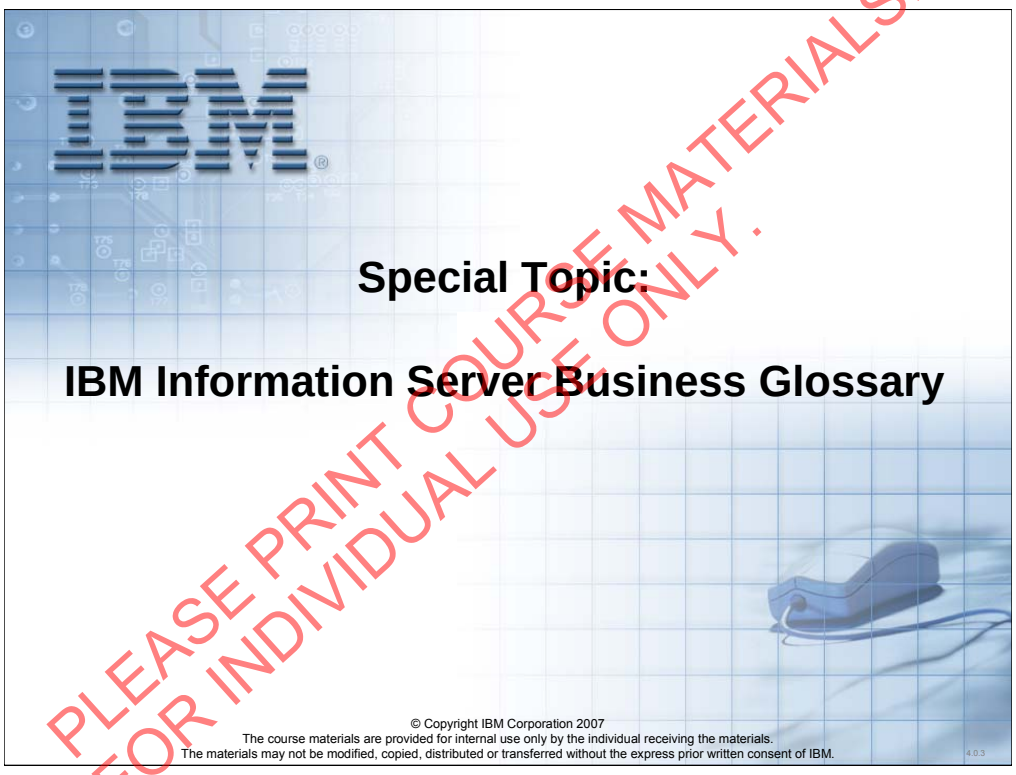
QSMT – expanded mode & running a QS8 job

- Prior to compiling, be sure to complete the following:
  - Provision any required rules to the server
  - Add ODBC connection information to any ODBC read or write stages appearing in the job
  - To complete the migration, perform the following for every Standardize, Survive, MNS and Waves stage that appears on the canvas:
    - Open the stage editor for the stage (e.g. by double-clicking it)
    - Click ok
- Once the above tasks are completed, compile and run as you would any other job

QSMT – expanded mode job results

- When running with a single node, output should be similar to that produced by QS 7.5 in stream mode
- When running with multiple nodes, output should be similar to that produced by QS 7.5 in PX mode when run with the same number of nodes

Note: If multiple 7.5 stages append to the same output, there are likely to be differences, as there is no guarantee in QS8 as to the order in which records will be added to the output file or stream in this case.



Agenda

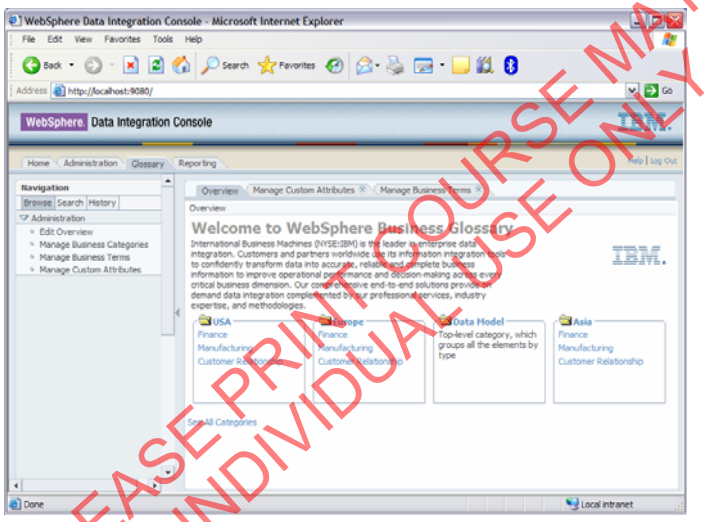
- Discussion of Business Metadata
  - **Introduction and Benefits**
  - Metadata Primer
  - Theory & Tools
- Product Overview
  - Purpose
  - Features & Functions
- Exercises
  - Accessing the Glossary
  - Administration
  - Uploading Categories and Terms
  - Browsing and Searching
  - Business Category and Term Creation and Editing
  - Annotations
  - Custom Attributes

Unit objectives

After completing this unit, you should be able to:

- Understand and Identify sources of metadata
- Differentiate between business and technical metadata
- Understand metadata as an asset and potential liability
- Understand and apply features and functions of the IBM Information Server Business Glossary

Introducing the IBM Information Server Business Glossary



An Intuitive Web-Based Tool for Authoring, Managing, and Sharing Business Metadata

This is Business Glossary Interface. It is based on the Mozart framework of the IBM Information Server Platform. The Glossary shares the same repository as the other tools of the IBM Information Server suite.

IBM Business Glossary helps business users with the following tasks:

Developing a common vocabulary between business and technology

- A common vocabulary allows multiple users of data to share a common view of the meaning of data. Users can assign categories and terms to data that are meaningful in a business context, and create a hierarchy of categories for ease of browsing.

Providing data governance and stewardship

- Data assurance programs assign responsibility to business users (data stewards) for the management of data through its lifecycle.

Finding business information that is derived from metadata

- Metadata helps business users to understand the meaning of the data, its currency, its lineage, and who is responsible for defining and producing the data. If a business user wants to know the definition of a term such as “corporate price,” the glossary will provide this insight.

Accessing metadata without complicated tooling and querying

- Metadata objects can be arranged in a hierarchical fashion to simplify browsing of the data objects.

Providing collaborative enrichment of business metadata

- Maintenance of business metadata is an ongoing process in which automated and manual data inputs evolve. Multiple business users can collaborate to add notes, annotations, categories, and synonyms to enrich business metadata.

For example, multiple systems may maintain tables of customer information, however the business may uncover a requirement for the concept of “high-value” customers. The business needs a way to define what a high value customer is, and how to recognize them (for example, a high-value customer is a customer with combined account balances over $10,000). IBM Information Server Business Glossary provides a tool for recording these definitions, and relating business concepts together into taxonomies. This records the business requirements in the same metadata foundation that the profiling and analysis process uses.

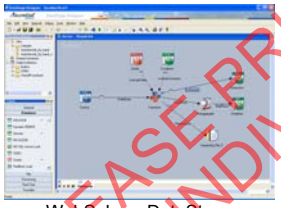
Business and technical users access the same repository

Database = DB2
Schema = NAACCT
Table = DLYTRANS
Column = ACCT_NO
data type = char(11)

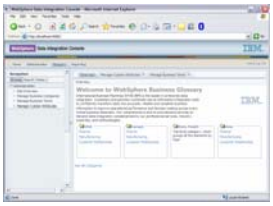
GL Account Number

The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Benefit: Create a common vocabulary between business & technical users



WebSphere DataStage



WebSphere Business Glossary

The Business Glossary shares the same repository as the other tools of the IBM Information Server.

Benefit: Collaborate by Sharing Experiences, Ideas, and Opinions

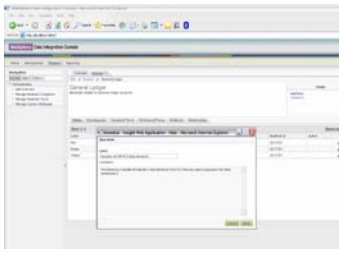
GL Organizational Unit

STEWARD: Controllers Office
FORMAT: X(7)
DEFINITION: A seven digit number designating the organizational unit to which this account belongs.

Author Standard Definitions

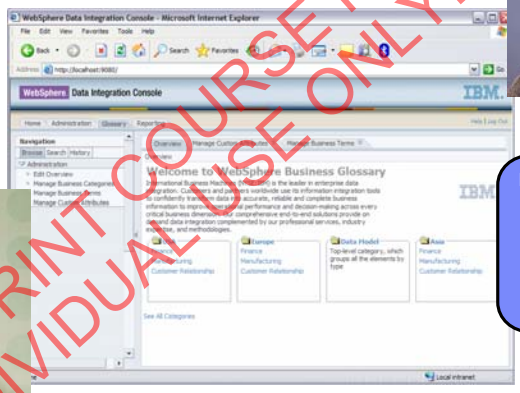
Annotate Any Object

I've noticed that the last two digits of the GL Organizational Unit, which indicate the sub-department, are often blank.



Benefit: Enable Quick Access to Critical Business Metadata

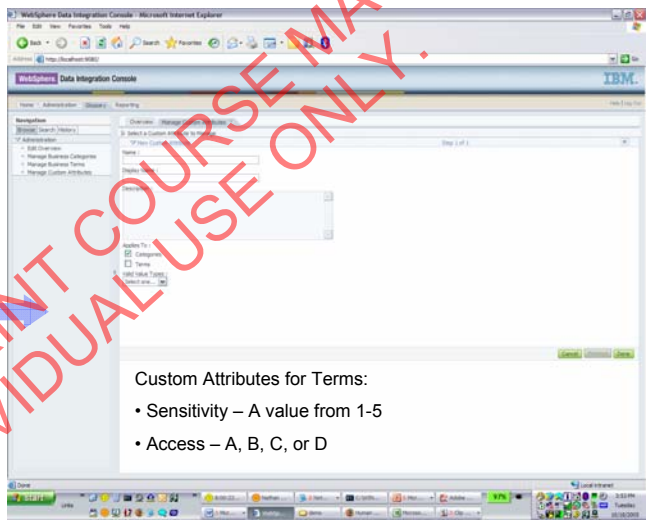
I didn't need any special training to use the business glossary. It's laid-out exactly as I expected.



I can set this up so that my people immediately see the most used business hierarchies

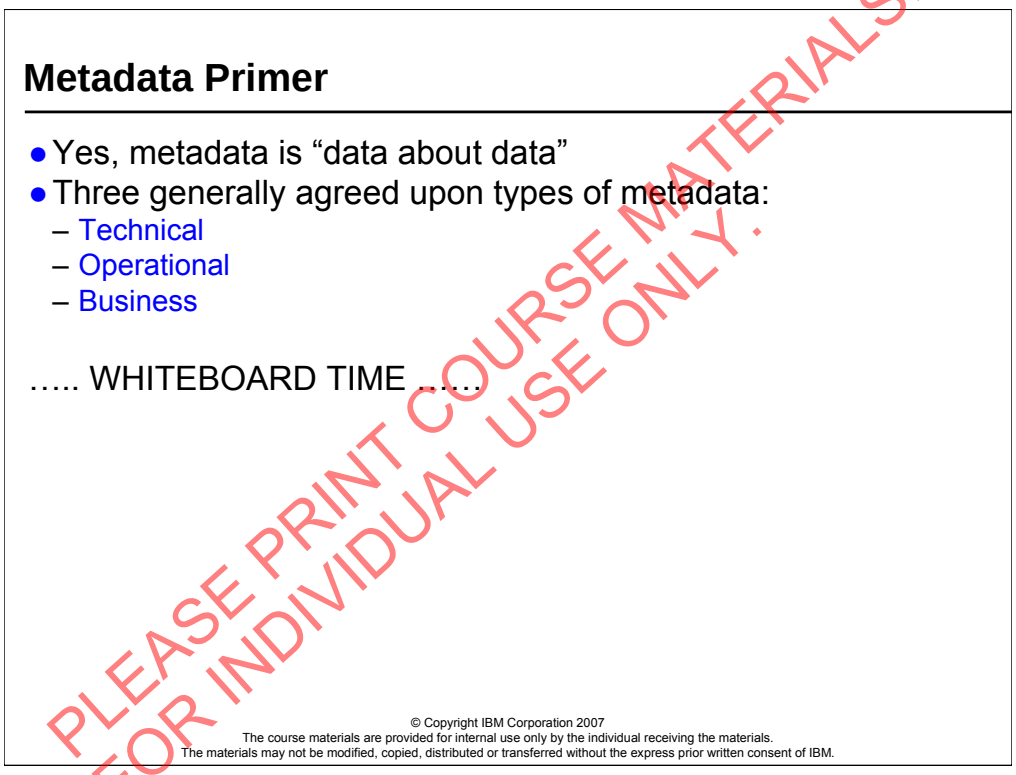
Benefit: Tailor the Application to Meet Your User's Specific Needs

Create custom attributes for which valid values must be set for every business term



Agenda

- Discussion of Business Metadata
  - Introduction and Benefits
  - **Metadata Primer**
  - Theory & Tools
- Product Overview
  - Purpose
  - Features & Functions
- Exercises
  - Accessing the Glossary
  - Administration
  - Uploading Categories and Terms
  - Browsing and Searching
  - Business Category and Term Creation and Editing
  - Annotations
  - Custom Attributes



Depending on your role or responsibility. The type of metadata of interest to you will vary. Consider the different types of metadata in your organization and enterprise. Examine the different uses of this metadata. For example, how a business user, data modeler or ETL developer might view an attribute such as account number. The business user might be interested in how this account number is traceable across operating units and is interested in the business process used for this traceability. The ETL developer having a different job function, might only be interested in the technical aspects of this attribute such as data type and potential growth of the field.

Business Metadata

- Documents the business meaning of data & related technology assets
- In the language of the business, independent of technology
- Used to
  - define a shared meaning of data
  - establish responsibility, accountability, and traceability
  - govern access
  - share insights & experiences among users
- Must be managed by those that understand the meaning and importance of the information assets to the business
- Better aligns the efforts of IT with the goals of the business

Business metadata answers questions like, “This report shows revenue; what kind of revenue? What is meant by revenue? What calculations went into the determination of revenue?” And questions like this: “How does this division of the enterprise calculate revenue? Is it the same way that our division calculates it?” In the past, much business metadata was stored in applications or stored procedures and triggers. Many technical tools do not support business rules; they support business rules translated into code. ETL tools, databases and data quality tools all show code, not English language descriptions. As stated above, business rules are supposed to be expressed in the language of the business.

Technical Metadata

- *Technical* metadata supports *software* efforts
- examples
  - Table and column definitions in a database schema
  - Mappings
  - Source and target system connection information
  - Stored procedures and triggers
  - Code
- Typically not in the language of the business

How Metadata Varies

GL Account Number

The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Business Metadata is about business meaning.

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = ACCT_NO

data type = char(11)

Not about technical details.

An illustration of two different user needs for same the term. You can see that they are viewed very differently according to the business role of the specific user.

The Ever Increasing Importance and Challenges of Metadata

- Understanding the meaning of information is important if you intend to use it to make decisions (obviously)
- Locating the information you need is becoming increasingly difficult in this world of ever increasing volumes of information
- Ditto for “directing” information, i.e. “I want other people to know this..”
- Corporate compliance issues are increasing the importance of understanding the business meaning of data across the enterprise
- Storing metadata in applications is no longer sufficient

The recent interest in semantics & ontology technology is being driven by this glut of information and the need for new technology to enable navigation through it.

Common Business Issues

- **Too much information and not knowing what's important**
  - Not using demand signals to drive supply chain
  - Not using customer analysis to tailor marketing and sales
  - Not leveraging valuable unstructured information
- **Multiple versions of the truth**
  - Problems managing customer, product and partner interactions
  - Regulatory compliance inhibited by poor transparency
- **Lack of trusted information**
  - Incomplete, out-of-date, inaccurate, misinterpreted data
  - Difficult to understand or control how information is used
- **Lack of agility**
  - Inability to take advantage of opportunities for innovation
  - Escalating costs due to inflexible systems and changing needs

Key Point: Customers tell us the inhibitors come down to four primary issues.

When we drill down one more layer on our conversations with customers, we find that many of them are facing the same kinds of problems.

Often they tell us that they **need to do a better job leveraging information.** This comes directly from the CEO survey mentioned on the previous slide. This problem manifests itself in many ways – sometimes companies complain about "information complexity", or a "deluge of information", but what that really means is that there is a great deal of valuable information locked away in various databases and systems throughout the business, but the organization has no easy way of using this information to improve the business, to compete more effectively, or to innovate. Often we see retail companies, for example, failing to use demand signals from their stores effectively to drive their supply chains. Across all industries, it is common to find that organizations are not using customer analysis to tailor their marketing and sales activities. In other cases, entire classes of information are being ignored, like unstructured information from emails and documents, simply because they are too difficult and expensive to deal with.

Another point we often hear is that organizations have **multiple versions of the truth** across their systems. This prevents them from being able to completely understand their customers, and tailor their interactions accordingly. It also leads to supply chain collaboration problems, since suppliers and customers have differing concepts and definitions of products. It also causes difficulties when trying to comply with information-centric regulations like Sarbanes-Oxley or Basel II, that require definitive information with associated proof.

The third thing we hear about is **trust** and **control**. Organizations don't have trust in their information because the quality cannot be assured, and the source of the information is often uncertain. At the same time, companies want to control who has access to information, understand how It is being used, and govern sensitive information throughout its lifecycle.

And lastly, organizations tell us that information inflexibility inhibits their **ability to respond quickly to change**. They can't take advantage of new opportunities for innovation, and their costs of maintaining IT systems continuously escalate as the business demands change from systems that weren't build for change.

What Does It All Mean?

- Business metadata defines the semantics of artifacts that are created & managed by other IT applications
- In other words… the meaning of data & supporting technologies are determined by business users
- A new technology that focuses on creating a **reusable** common business vocabulary is critical to the effective use of these assets

We need a tool that enables business users to document the business meaning of these assets and attach to the artifacts that are collected & created by the other tools involved in information management.

Enterprise Objective:

Document & share critical business aspects of data integration artifacts – especially the data itself.

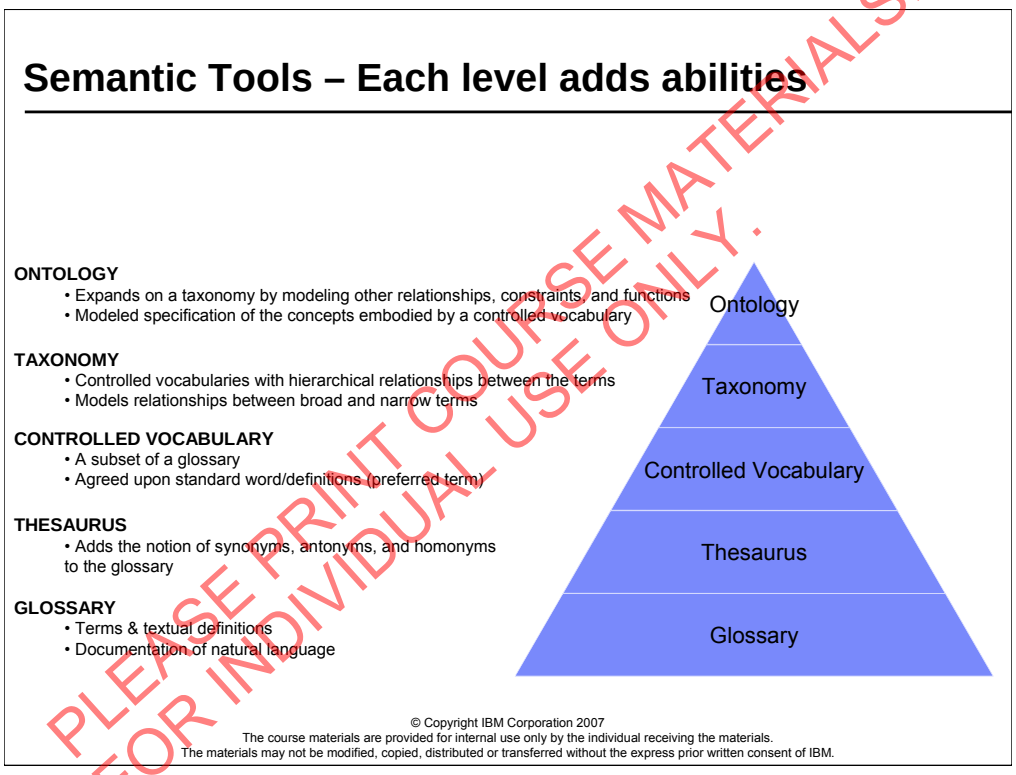
- There is a core set of information that all enterprises require
  - Standard names & definitions for date items (terms)
  - Organized as hierarchies
  - With descriptions, examples, abbreviations, and stewardship information
- Additionally, each customer has a unique set of business metadata that is critical to their organization
  - Calculations
  - Policies
  - Validation
  - Authority
  - Security
  - Sensitivity

GL Account Number:
The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

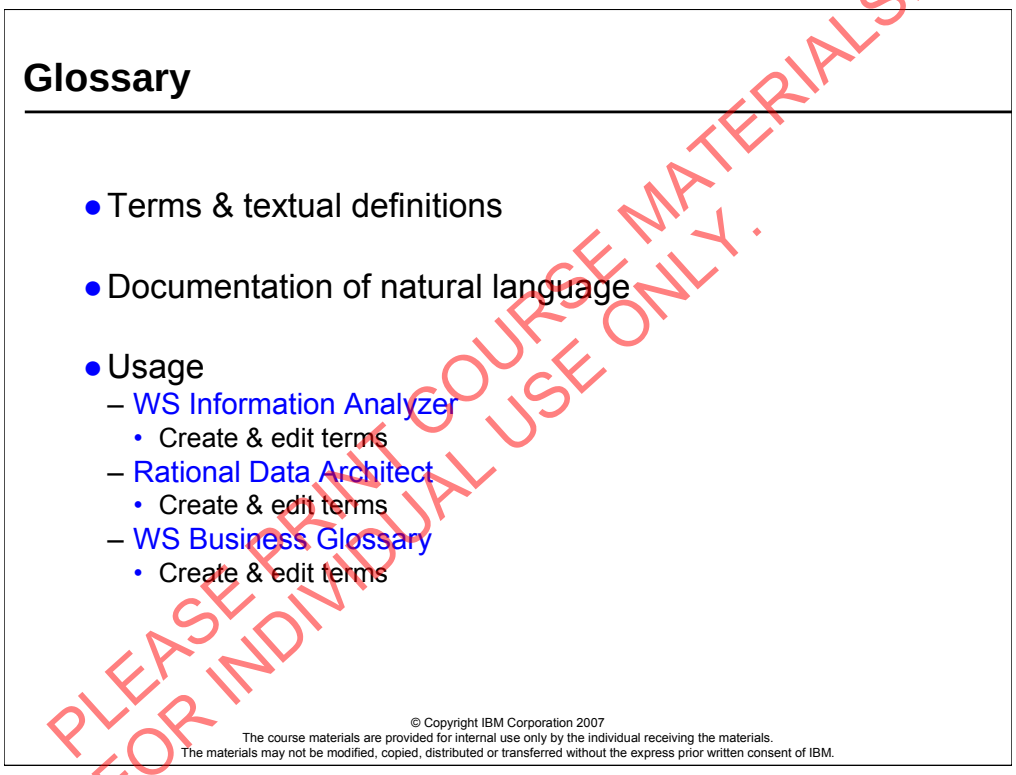
Data Sensitivity
A number 1-5, indicating the sensitivity of the data. Sensitivity is a subjective measure of the impact of said data being released to unauthorized consumers.

Agenda

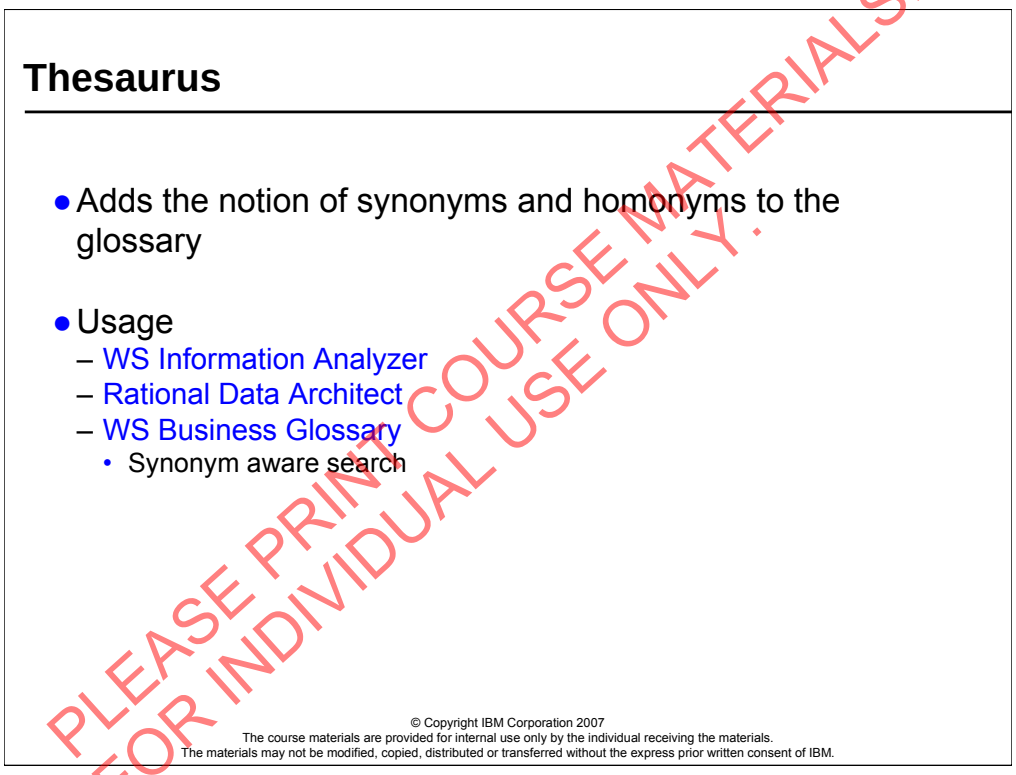
- Discussion of Business Metadata
  - Introduction
  - Metadata Primer
  - **Theory & Tools**
- Product Overview
  - Purpose
  - Features & Functions
  - Benefits
- Exercises
  - Accessing the Glossary
  - Administration
  - Uploading Categories and Terms
  - Browsing and Searching
  - Business Category and Term Creation and Editing
  - Annotations
  - Custom Attributes



The levels and evolution of information cataloging. Each level of this pyramid adds a additional capabilities to the basic storage of business terms. By leveraging these “tools” we can bring more meaning and utility to our metadata assets.

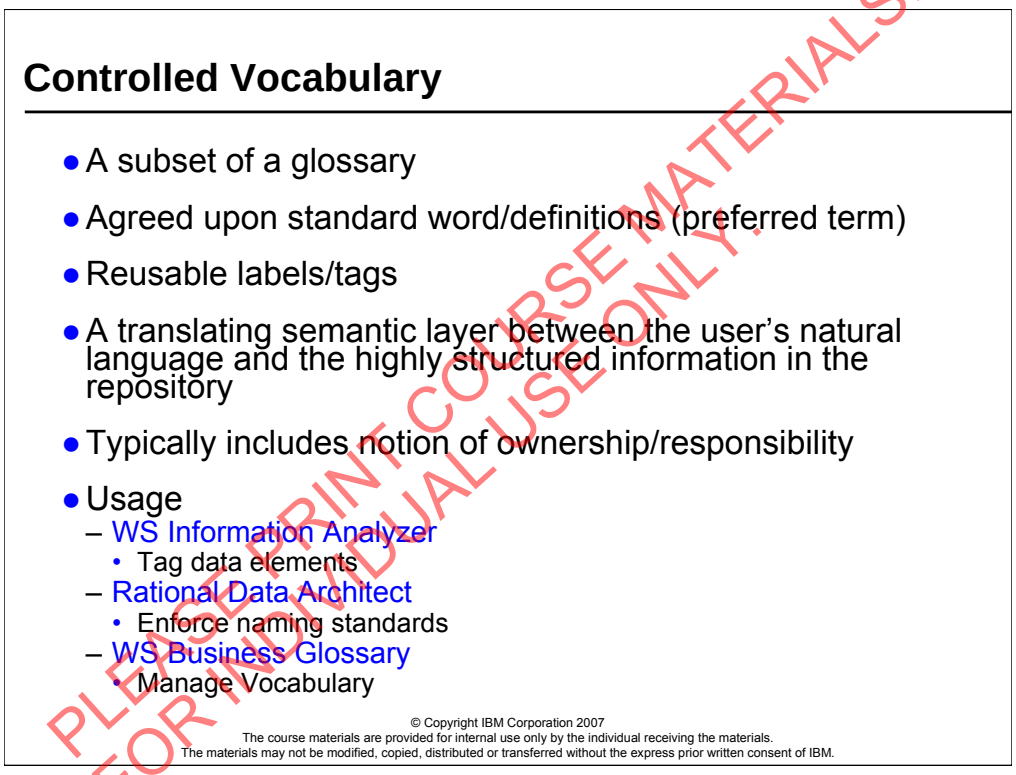


A glossary is our simplest grouping of business terms. Most of us are familiar with text book glossaries and the indexes that typically accompany them. This combination is not practical with the scale of data of even the smallest database.



If your metadata repository includes some different terms that mean the same thing, you can designate such terms as synonyms. If two terms are not synonyms, but are related in some other way that is important, you can designate them as related terms. You can specify which term of a group of terms is the preferred term, and which terms to replace with other terms. You can also specify standard abbreviations of the term.

If your metadata repository includes some like sounding terms, but with very different meanings, this can be identified within the glossary.



Takes the Guess Work out of Searching
A controlled vocabulary makes a database easier to search. Since we have many different ways of describing concepts, drawing all of these terms together under a single word or phrase in a database makes searching the database more efficient as it eliminates guess work. However, arriving at this efficiency requires consistency on the part of the individual indexing the database and the use of pre-determined terms.

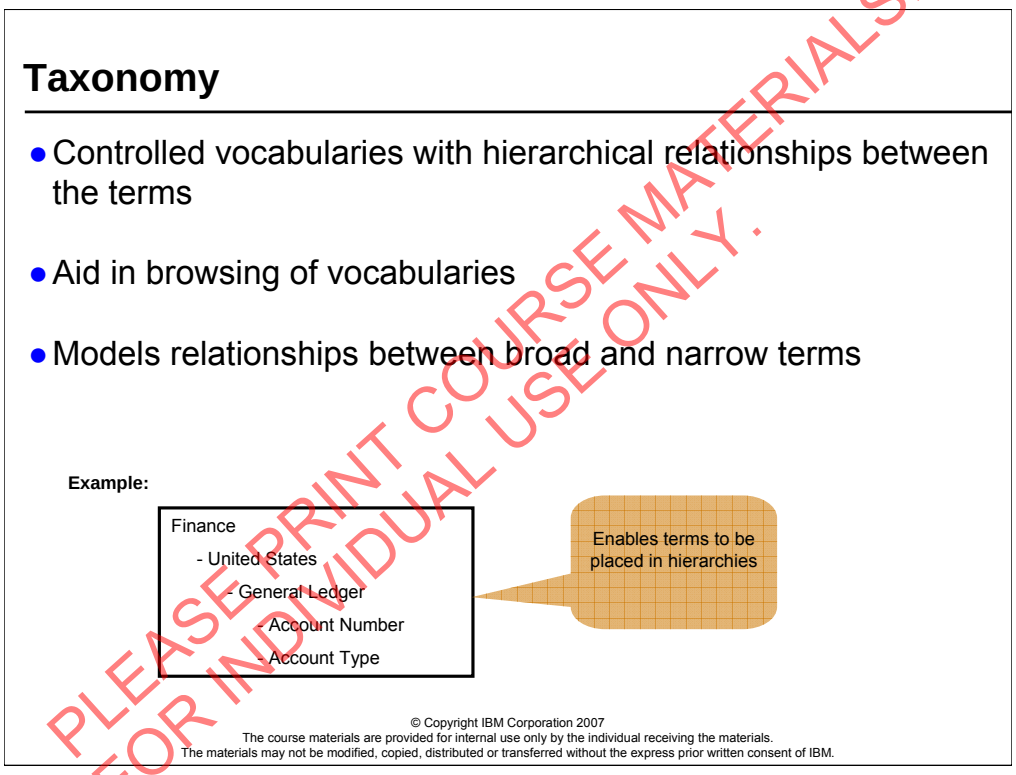
A Familiar Concept
It's likely you are already familiar with the concept of controlled vocabulary. Phonebook Yellow Page listings are arranged using controlled vocabulary. For example, a search for "Car Dealers" leads you to a note to "see Automobile Dealers." At a basic level, this is how a controlled vocabulary system works.

One Search is All it Takes
Conducting a search in a database that uses controlled vocabulary or indexing terms is efficient and precise. The biggest advantage to controlled vocabulary is that once you do find the correct term, most of the information you need is grouped together in one place, saving you the time of having to search under all of the other synonyms for that term.

Testimonials for a Common Business Vocabulary

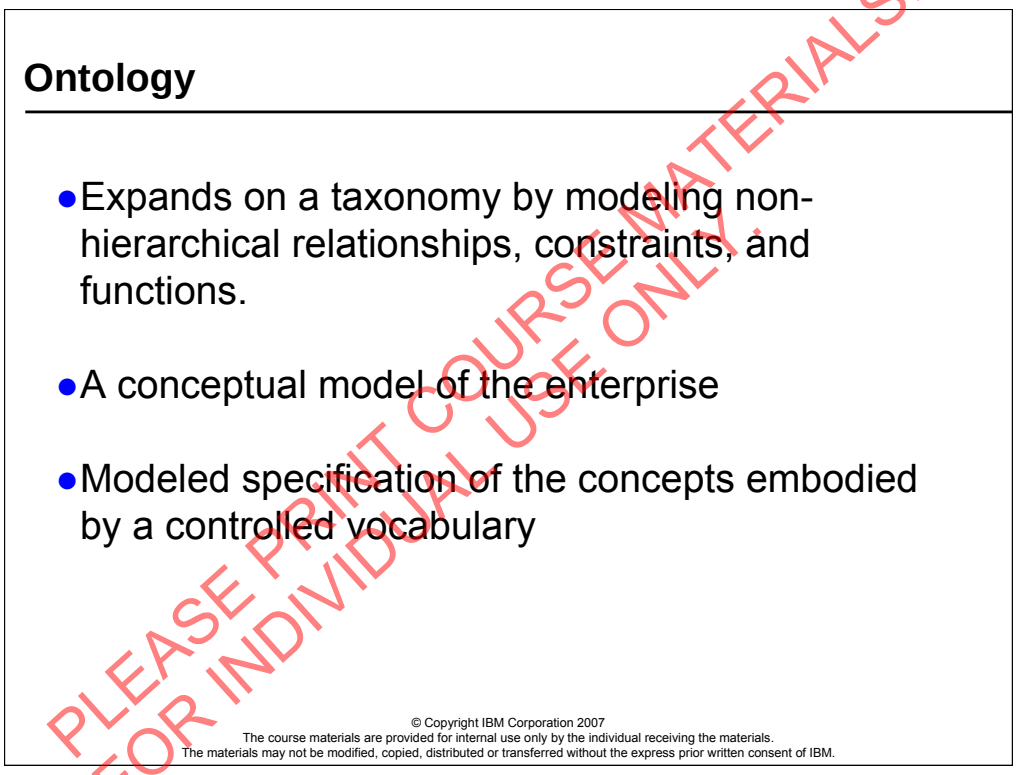
- "Having a common business language can minimize the time and effort (of business process integration), and make these critical business processes much more efficient" (Gartner)
- "Having a common business language helps to encourage transparency, quality of information, privacy and security of information" (Gartner)
- "The idea of using a common language to… decouple applications from data – is another potential impact on the IT stack. " (Gartner)

Common or Controlled Vocabulary are often referred to synonymously.



The classification, or categorization, of things. For example, a Web taxonomy would classify all the sites on the Web into a hierarchy for searching purposes. This comes from the Greek words "taxis" and "nomos," which mean "division" and "law."

In theory, the development of a good taxonomy takes into account the importance of separating elements of a group (taxon) into subgroups (taxa) that are mutually exclusive, unambiguous, and taken together, include all possibilities. In practice, a good taxonomy should be simple, easy to remember, and easy to use. One of the best known taxonomies is the one devised by the Swedish scientist, Carl Linnaeus, whose classification for biology is still widely used (with modifications). In Web portal design, taxonomies are often created to describe categories and subcategories of topics found on the Web site.



An **ontology** is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them.

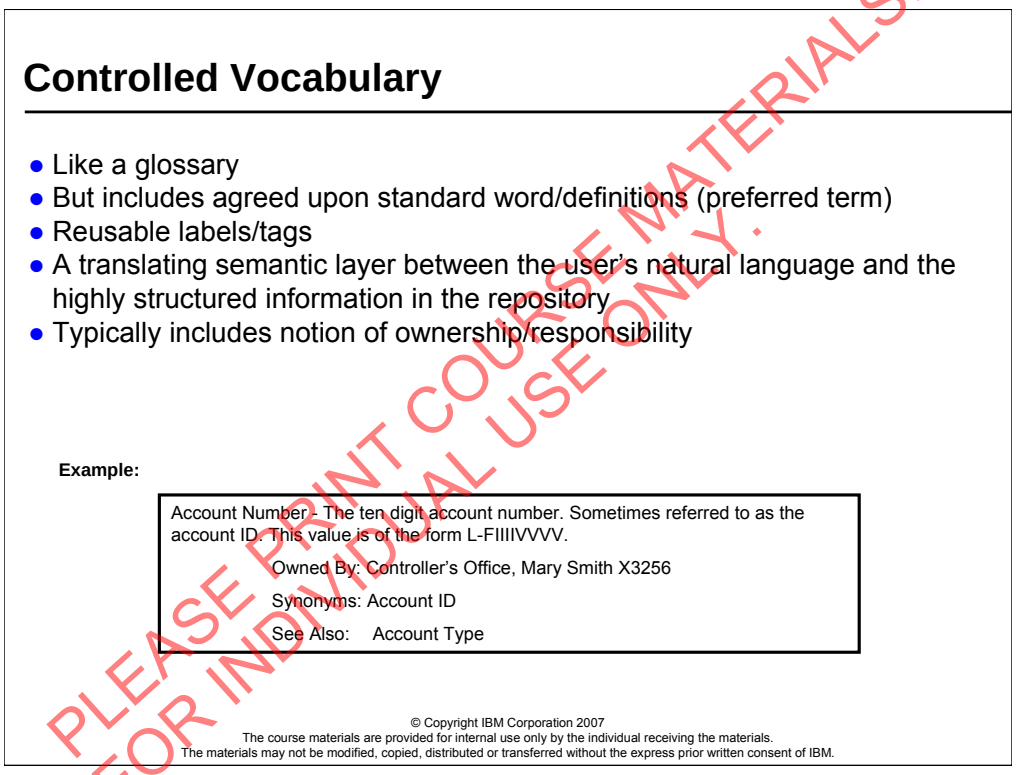
Ontologies are used in artificial intelligence, the semantic web, software engineering and information architecture as a form of knowledge representation about the world or some part of it.

The periodic table of elements is a classic example of a widely used taxonomy. It becomes an ontology when it is combined with the following assertions and constants.

- A molecule is the smallest quantity of a compound composed of chemically bonded elements
- A chemical reaction occurs when reactants (elements and/or molecules) combine to produce products (elements and/or molecules) of different composition
- Mass is conserved in a chemical reaction
- Elements cannot be transmuted in a chemical reaction
- A Mole of any compound is the product of the Avogadro number and the sum of the atomic masses of its constituent elements

This classic chemistry ontology has withstood the attacks of scientist and abuse of school children for 200 years. It allows chemists to test the plausibility of any possible chemical reaction and predict the quantities of reaction products. The system is not perfect it cannot predict if reactions are thermodynamically likely, and it cannot predict some element properties, for example why Mercury is a liquid, but it can rule out many implausible chemical reactions. Most spectacularly it has been used to infer the existence of undiscovered elements and compounds.

From this type of example we can could perhaps relate it to the data model of our enterprise. What would happen if took a customer or finance record from different operating units. From examining the relationships and business rules of the contained terms in these records, we can determine the impact or possibility of merging these records.



A **Controlled vocabulary** is a carefully selected list of words and phrases, which are used to tag units of information so that they may be more easily retrieved by a search

Agenda

- Discussion of Business Metadata
  - Introduction and Benefits
  - Metadata Primer
  - Theory & Tools
- Product Overview
  - **Purpose**
  - Features & Functions
- Exercises
  - Accessing the Glossary
  - Administration
  - Uploading Categories and Terms
  - Browsing and Searching
  - Business Category and Term Creation and Editing
  - Annotations
  - Custom Attributes

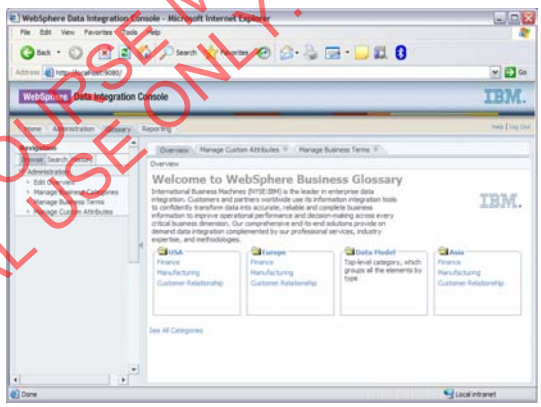
Unit objectives

After completing this unit, you should be able to:

- Understand and the apply the following features and functionality of the IBM Information Server Business Glossary
  - Categories
  - Terms
  - XML upload capability
  - Synonyms
  - Custom attributes
  - Administration
  - Stewardship
  - Modification of the Welcome screen

Introducing IBM Information Server Business Glossary

- A web-based tool that for business users that enables
  - Creation & management of a controlled vocabulary
  - Creation & management of a business taxonomy
  - Collaborative authoring of business metadata
- A reference for learning about the information assets of the enterprise
  - Meaning
  - Usage
  - Ownership/Responsibility



- Organized according to the business hierarchies defined by the business taxonomy

Managing business metadata effectively can ensure that the same data “language” applies throughout the organization. IBM Information Server Business Glossary gives business users the tools they need to author and own business metadata.

For example, one department refers to “revenues,” another to “sales.” Are they talking about the same activity? One subsidiary unit talks about “customers,” another about “users” or “clients.” Are these different classifications or different terms for the same classification?

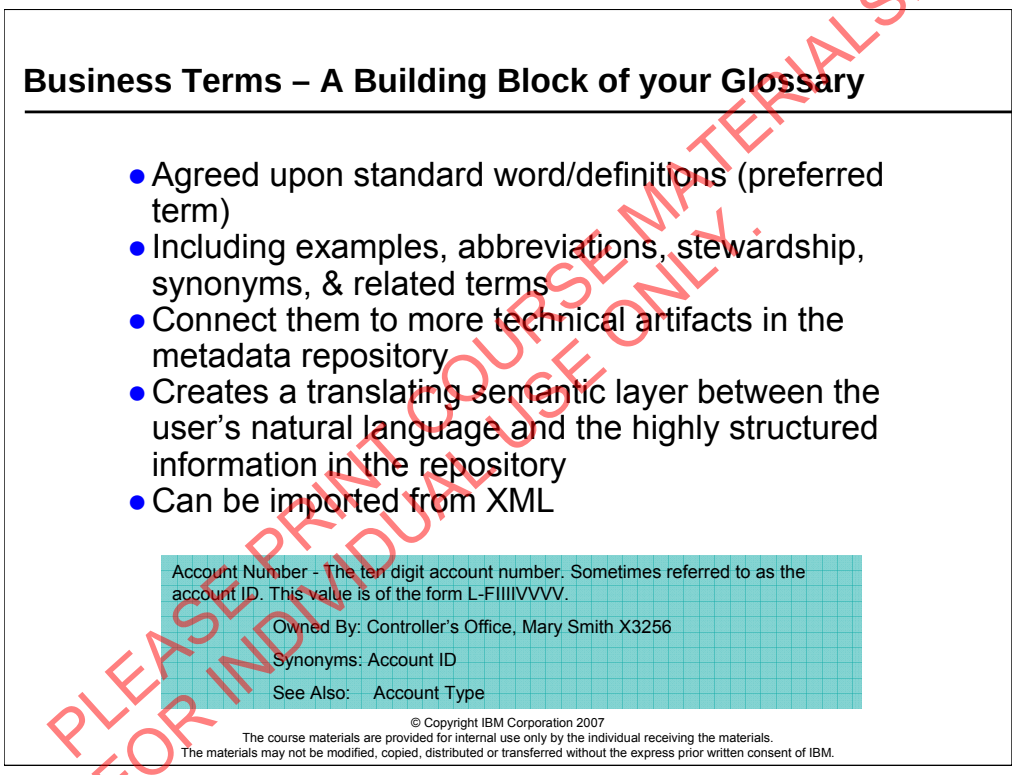
IBM Information Server Business Glossary provides business users with a Web-based tool for creating and managing standard definitions of business concepts, called a *controlled vocabulary*. It also simplifies the building of a business-oriented classification system and the collaborative authoring of business metadata.

The tool simplifies the task of managing, browsing, and customizing the broad variety of metadata that is stored in the repository of IBM Metadata Server, metadata that includes details about tables, columns, models, schemas, operations, and other components of the data integration process.

The tool divides metadata into categories, each of which contains terms. You can use terms to classify other objects in the metadata repository based on the needs of your business. You can also designate users or groups as stewards for metadata objects.

Agenda

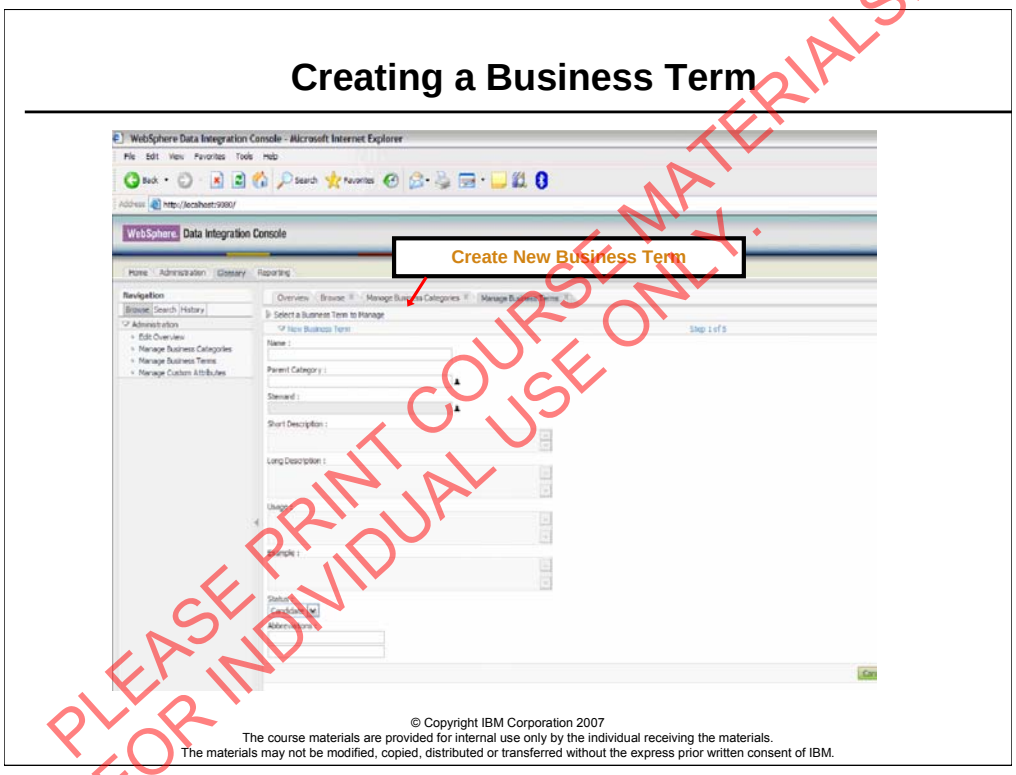
- Discussion of Business Metadata
  - Introduction and Benefits
  - Metadata Primer
  - Theory & Tools
- Product Overview
  - Purpose
  - **Features & Functions**
- Exercises
  - Accessing the Glossary
  - Administration
  - Uploading Categories and Terms
  - Browsing and Searching
  - Business Category and Term Creation and Editing
  - Annotations
  - Custom Attributes



Administrators and authors can use terms to classify objects in the metadata repository.

A term is a word or phrase that can be used to classify and group objects in the metadata repository. For example, you might use the term Africa Sales to classify some of the tables and columns in the metadata repository, and the term Europe Sales to classify other tables and columns.

If your metadata repository includes some different terms that mean the same thing, you can designate such terms as synonyms. If two terms are not synonyms, but are related in some other way that is important, you can designate them as related terms. You can specify which term of a group of terms is the preferred term, and which terms to replace with other terms. You can also specify standard abbreviations of the term.



Term properties

You can specify the following properties of a term:

Name

Term names must be unique. If you attempt to create a term with the same name as an existing term, you are prompted to use a different name.

Names of terms must start and end with a character that is not a space. Names cannot contain any of the following characters:

. (period)

, (comma)

; (semicolon)

% (percentage sign)

" (quotation marks)

Parent Category

The category that contains the term. A term must have one and only one parent category.

Short Description

Optional. Short descriptions are important because they can help uniquely identify a category in a list of other terms with similar names. The text should be no longer than one or two lines. Short descriptions are used in many searches and are displayed in lists of objects.

Long Description

Optional. They can contain additional information that might be of interest to users who are browsing the term or administrators and authors who are setting its relationships to other objects.

Usage

Optional. Information about how to use the term, and any business rules that govern its use.

Example

Optional. An example of how the term is used, or a typical sample value.

Status

The approval status of the term within the organization. Status has one of the following values:

Candidate

The default value for new terms.

Accepted

Accepted by an administrator for general use.

Standard

Considered the standard for definitions of its type.

Deprecated

Should no longer be used.

Type

The classification of a term based on its use for metadata naming standardization. Type has one of the following values:

None

The type has not been declared. This is the default value.

Primary

The term describes a major enterprise concept such as a customer or an employee. Primary terms are usually nouns or noun phrases that form the basis for naming objects in data models.

Secondary

The term identifies a secondary distinguishing characteristic of a business concept, such as an identification number. Secondary terms are usually nouns or noun phrases that form the basis for naming an attribute of an object.

Is Modifier

Describes whether or not the primary purpose of the term is to provide descriptive information about an object. Is Modifier has the following possible values:

Yes

The primary purpose of the term is to provide descriptive information about an object.

No

The primary purpose of the term is to identify distinguishing characteristics of an object. No is the default value.

Preferred Synonym

The term is the preferred term in a group of synonym terms. Terms with the Deprecated status cannot be preferred terms.

Abbreviations

Optional. One or two standard abbreviations of the term.

Term relationships

You can specify that terms have relationships to the following types of objects:

Steward

The person or group that is responsible for the term. A term can have only one steward.

Related Terms

Terms that are related in some way to the term in question. This relationship can be used for *see also* relationships to terms that are similar but not identical. The relationship is not symmetrical. If you specify that term A has term B as a related term, that does not imply that term B has term A as a related term. A term can have multiple related terms.

Synonyms

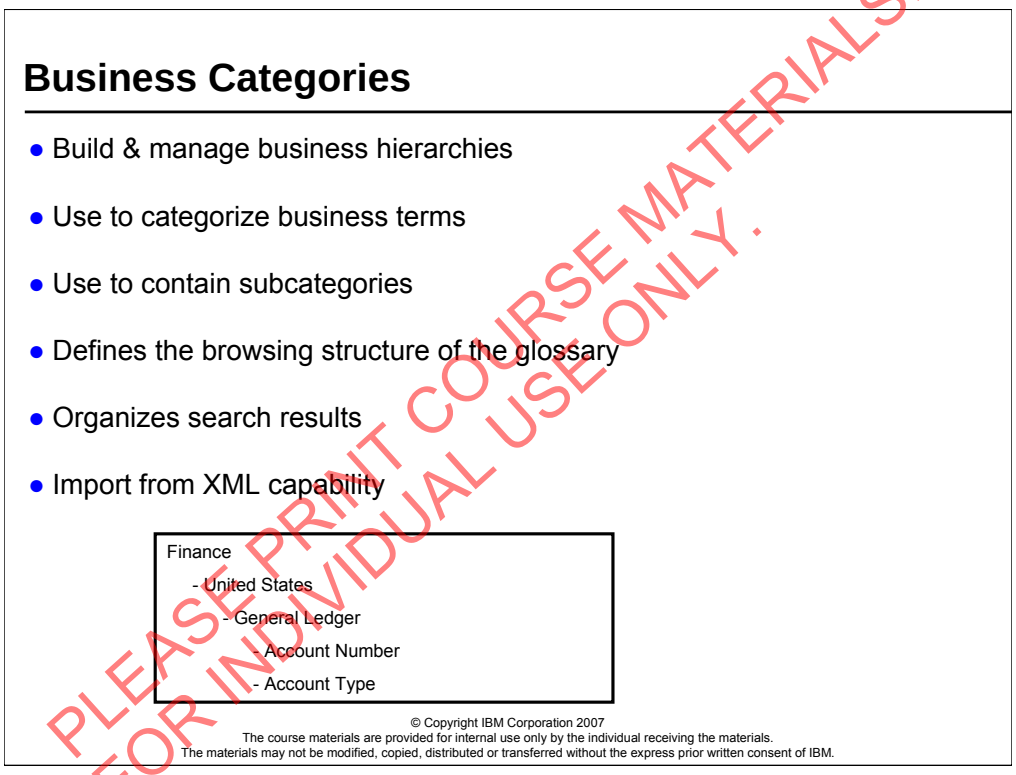
Terms that have the same meaning. A term can have multiple synonym terms. The relationship is symmetrical and transitive. If term A is a synonym of term B, and term B is a synonym of term C, each term is a synonym of the others.

Classified objects

Objects that the term classifies. A term can classify multiple objects in the repository. An object can be classified by multiple terms.

Custom attributes

If an administrator created custom attributes that apply to terms in general, you can edit values for these attributes when you create or edit a particular term.



Administrators and authors create, edit, and upload categories to provide the structure for the glossary.

A category functions much like a folder or directory. A category can contain other categories and terms. These containment relationships provide the logical structure for the glossary, so that users can easily browse and understand the metadata in the metadata repository.

In addition, a category can reference terms that it does not contain. The reference relationship is essentially a pointer to a term that is contained by another category, but that has relevance to the referencing category.

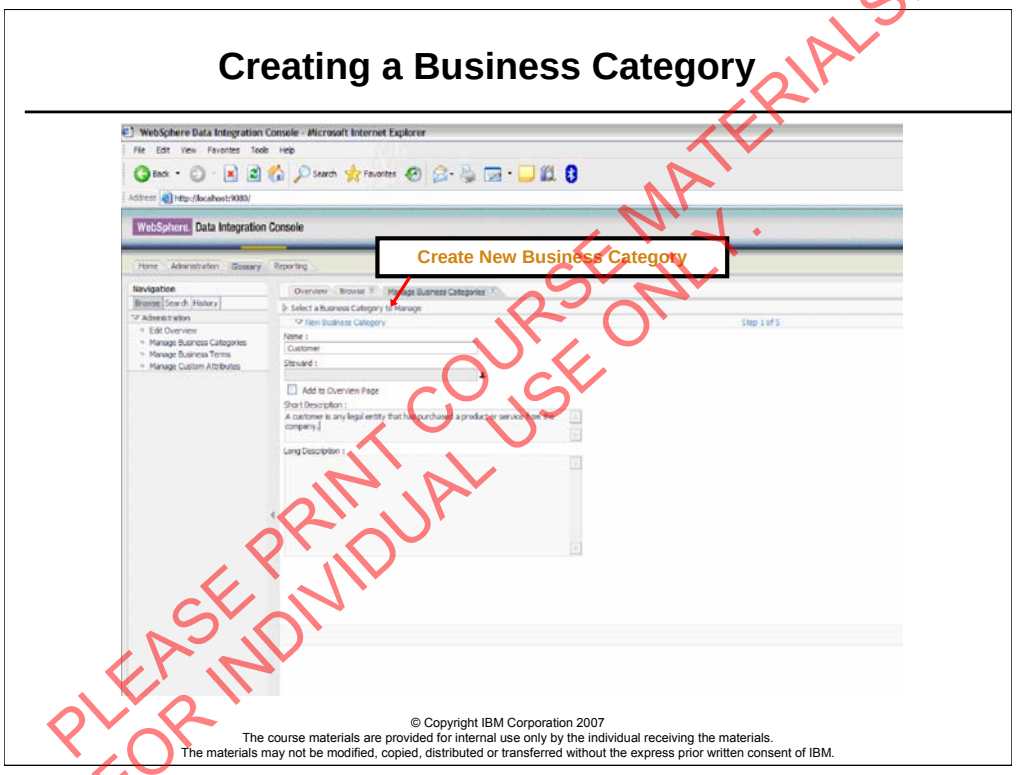
When creating or editing a category, you can do the following actions:

Specify category properties.

Specify category relationships, including subcategories, contained terms, and referenced terms.

Edit values for any custom attributes that apply to the category.

Administrators and authors can also upload files of categories and terms to the metadata repository, and then specify additional properties and relationships.



Category properties

You can specify the following properties for a category:

Name

- Categories can have identical names only if they have different parent categories. For example, categories named Asia and Europe can each contain a different category named Sales. But the category Asia can contain only one category named Sales.
- Names of categories must start and end with a character that is not a space. Names cannot contain any of the following characters:

  . (period)

  , (comma)

  ; (semicolon)

  % (percentage sign)

  " (quotation marks)

Short Description

- Optional. Short descriptions are important because they can help uniquely identify a category in a list of other categories with identical or similar names. The text should be no longer than one or two lines. Short descriptions are used in many searches and are displayed in lists of objects.

Long Description

- Optional. They can contain additional information that might be of interest to users who are browsing the category or administrators and authors who are setting its relationships to other objects.

Category relationships

You can specify that a category has relationships to the following types of objects:

Parent category

- Optional. A category can have only one parent category.

Steward

- Optional. The steward is the person or group that is responsible for the category. A category can have only one steward.

Subcategories

- Optional. A category can have multiple subcategories. A category is the parent category for its subcategories. These subcategories can also have subcategories, and can contain terms or reference terms.

Contained terms

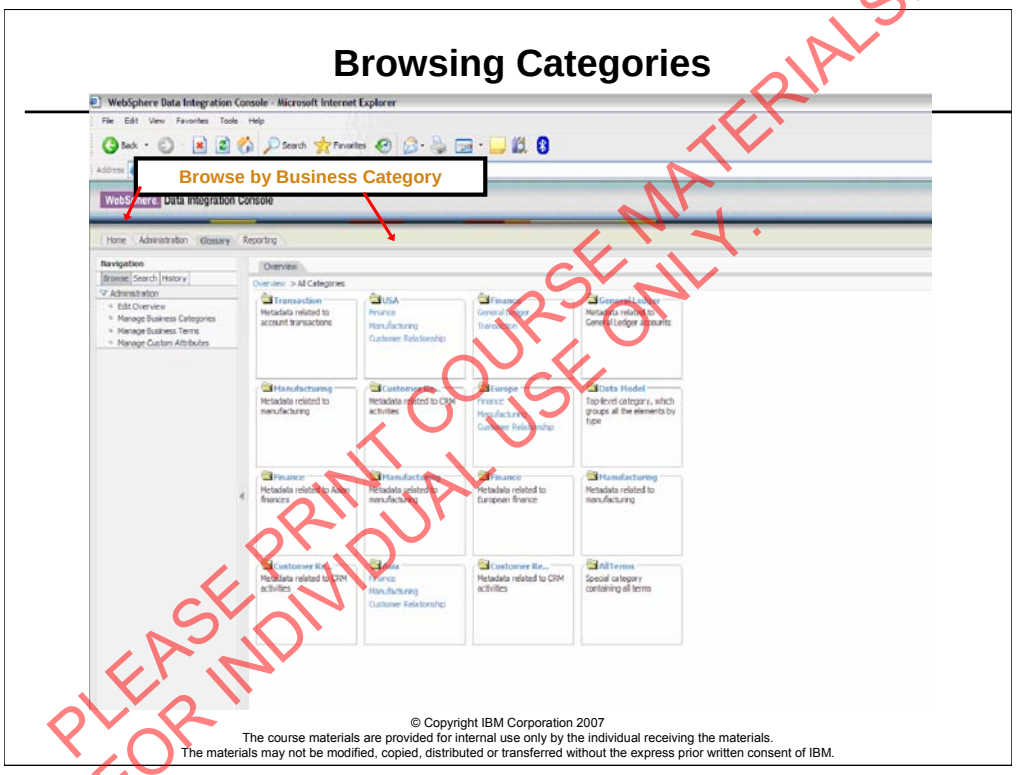
- Optional. A category can contain multiple terms. Each term must be contained by a single category.

Referenced terms

- Optional. A category can reference multiple terms. Each term can be referenced by multiple categories. You can use this relationship to provide pointers from a category to a term that is not contained by the category, but that has a relationship to the category.

Custom attributes

If you create custom attributes that apply to categories in general, you can edit values for these attributes when you create or edit a category.



You can browse the glossary structure to explore categories, terms, and objects in the repository of IBM Information Metadata Server.

You can start browsing the glossary from the Overview page, which displays the top-level categories that the glossary administrator has designated as most important for navigation in the metadata repository. You can also search for objects and select an object from the search results.

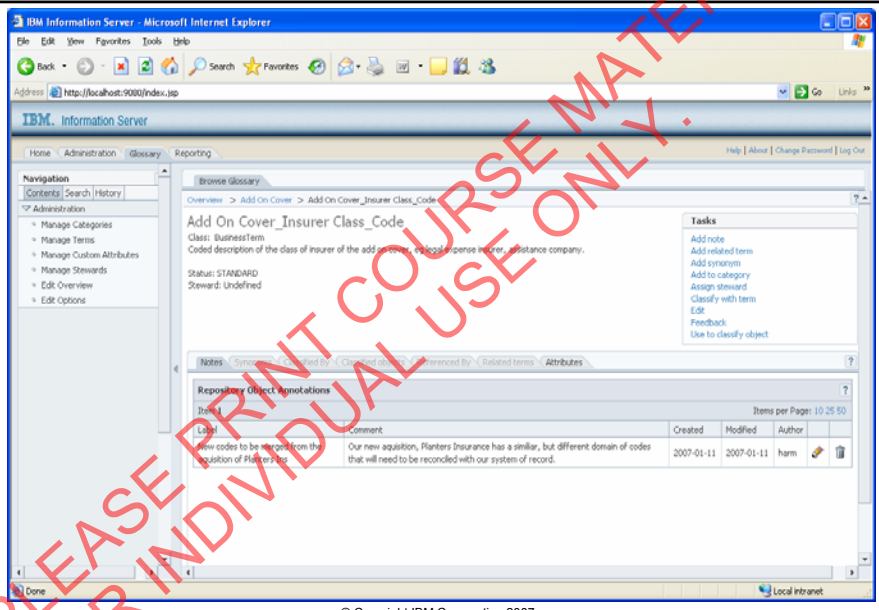
When you select an object, the browse page of the object is displayed on the Browse Glossary tab, which lists the name, class, steward and other important properties of the object. You can inspect the attributes of the object, browse its relationships to other objects, and send feedback to the administrator. Administrators and authors can add and edit notes about the object.

Collaborative Authoring of Business Metadata

- Add Notes/Annotations to any object in the glossary
- Enables business users to enrich the metadata associated with information assets
- Similar to the user feedback section on many online storefront sites
- Immediately consumable by other users
- Monitored by administrator
- Provide feedback to system administrator

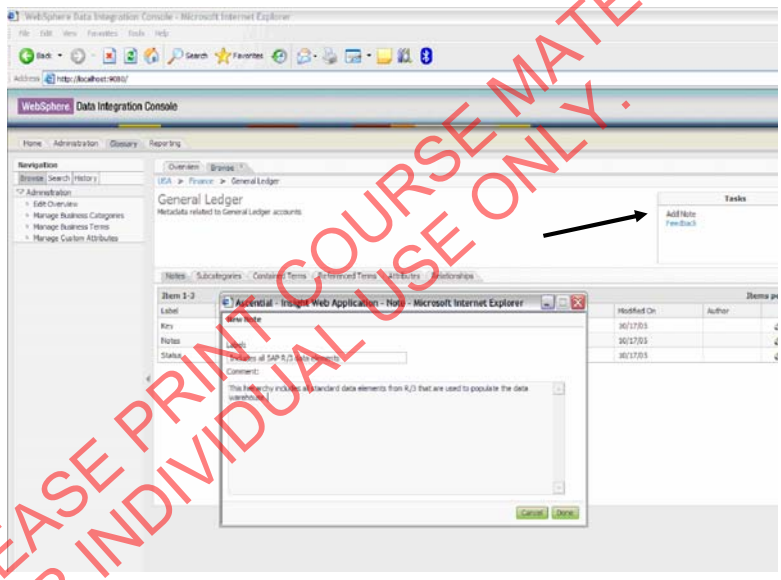
I've noticed that the last two digits of the GL Organizational Unit, which indicate the sub-department, are often blank.

Business User Annotations (Notes)



Notes are a mechanism to share observations and information about the metadata in the repository. If a note has been created for a category or a term. The Notes tab will be enabled. Selecting the Notes tab will a show the detail of the Note for that term or category.

Adding Notes



Adding, editing, and deleting notes

Administrators and authors can add, edit and delete notes on the browse page of any object.

Prerequisites: You must have the Business Glossary Administrator or Business Glossary Author role to add notes. Administrators can edit and delete all notes. Authors can edit and delete only the notes that they added. To add, edit, or delete notes:

Display the browse page of an object by any of these methods:

- Browsing from the Overview page
- Finding objects by using a simple search
- Finding objects with an advanced search
- Browsing the properties and relationships of objects

Add, edit, or delete the note: **OptionDescription*To add a note:***

- *In the **Tasks** list, click **Add Note**.*
- *In the New Note window, type a label and comment for the note and click **OK**.*

The note is added to the Notes tab.

To edit a note:

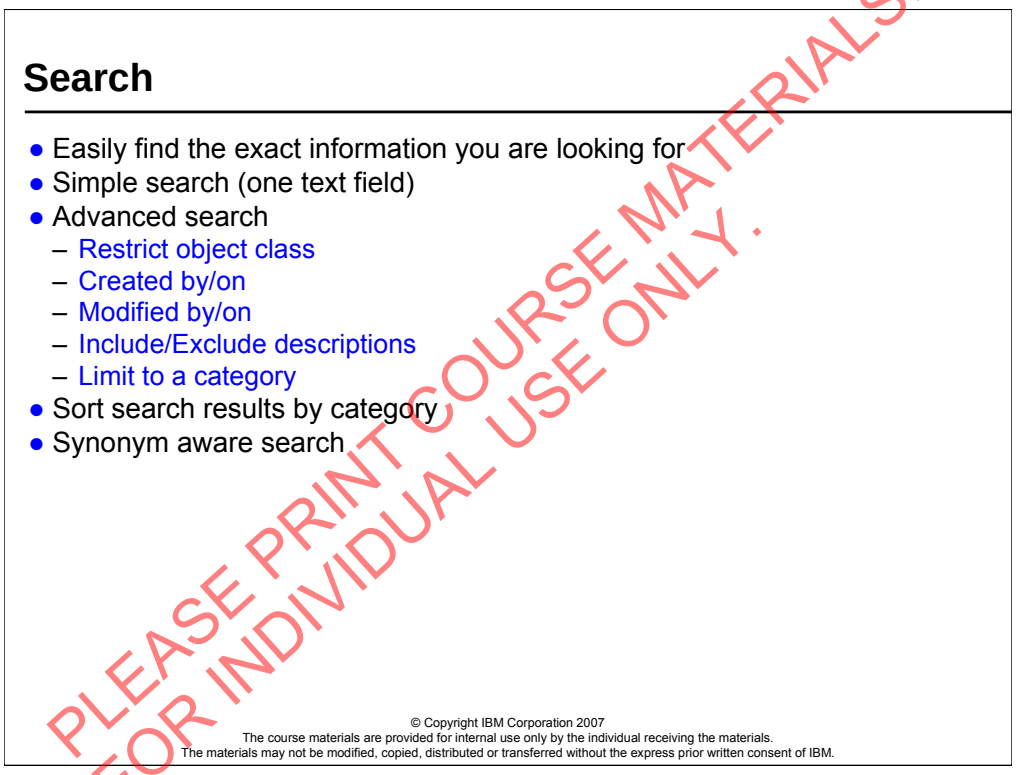
- *On the Notes tab, in the row that describes the note that you want to edit, click (edit note). The icon is not displayed if you do not have authority to edit the note.*
- *In the Edit Note window, type a label and comment for the note and click **OK**.*

To delete a note:

- *On the Notes tab, in the row that describes the note that you want to delete, click*

(delete note). The icon is not displayed if you do not have authority to delete the note.

- *Click **Yes** to confirm deletion.*



Searching for objects

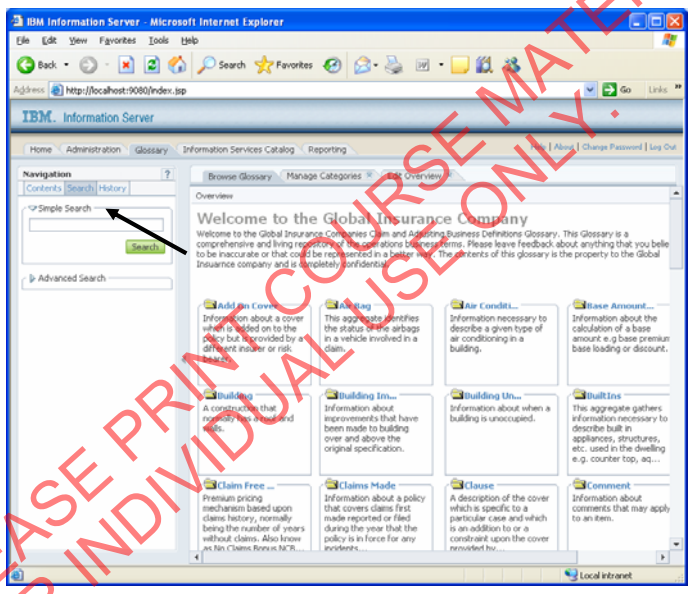
Using the search tool is often the quickest way to find an object in the repository of WebSphere® Metadata Server.

You can perform simple and advanced searches to find repository objects of all classes, including, but not limited to, categories, terms, tables, columns, job definitions, users, and groups. The more information that you can specify about the object that you are searching for, the faster the search results are returned.

The containment path of the object is displayed in the Path column in the search results so you can distinguish between objects with similar or identical names. For example, the containment path for a column might display the names of the containing table, schema, database, and host computer the column was imported from.

When you locate the object in the search results, you can click its name to display the browse page for the object. You can then inspect its attributes, browse its relationships to other objects, and send feedback to the administrator. Administrators and authors can add and edit notes about the object.

Simple Search



Finding objects by using a simple search

You can search for objects that are stored in the metadata repository based on the object name or description.

To find objects by using a simple search:

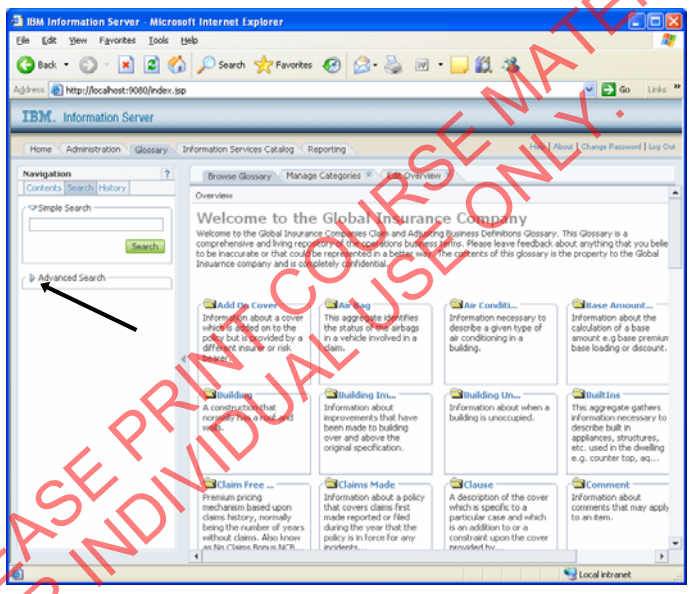
In the Navigation pane on the Glossary tab, select **Search** > **Simple Search**.

In the **Simple Search** field, type the search criteria. You can use all or part of a name or short description, or you can use multiple keywords from names or descriptions, separated by spaces or commas.

Click **Search**. The Search Results page displays a list of objects in the metadata repository whose names or short descriptions match the search string. If you typed multiple keywords, the list includes the objects whose names or short descriptions include all of the keywords.

In the list, click an object name to view the object, its relationships, and its attributes.

Advanced Search



Finding objects with an advanced search

You can use multiple criteria when you search for objects that are stored in the metadata repository.

To find objects with an advanced search:

In the Navigation pane on the Glossary tab, select **Search** > **Advanced Search**.

Specify the criteria for the search: **OptionDescription***To search by keyword:**In the **Keyword** field, type one or more keywords that represent words in the name, short description, or long description of the object. Use commas or spaces to separate multiple keywords. You can leave this field blank when you are searching by class or date.* ***To define the scope of the keyword search:****Select one of the following options:*

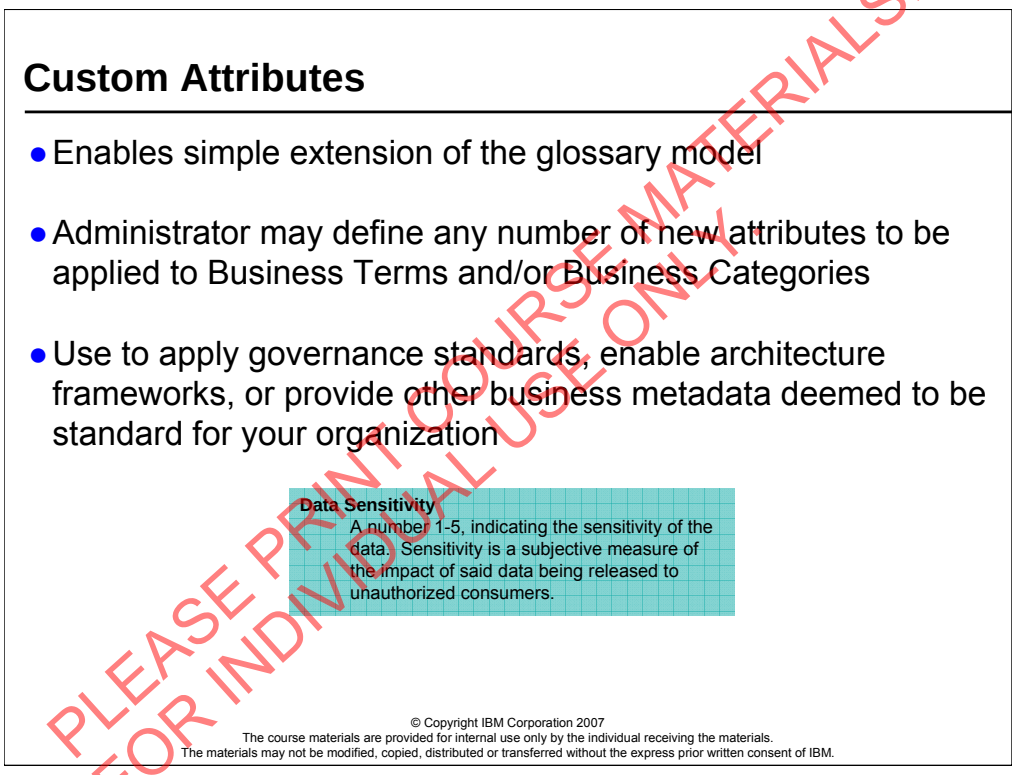
- ***All****. Finds objects where all keywords are present.*
- ***Any****. Finds objects where any of the keywords are present.*
- ***Exact****. Finds objects where the exact search string entered in the* ***Keyword*** *field is present.*

To search in the description field of objects:*Select* ***Search Descriptions****. Otherwise, only the name field is searched.****To display synonym terms of any terms returned by the search:****Select* ***Also Return Synonyms****.****To limit the search to a particular class of metadata object:*** *Select the class from the* ***Class*** *drop-down list. For example, to limit the search to categories, select*

BusinessCategory*. You can run this search without specifying a keyword.****To limit the search to objects that were created by a particular user whose identity is defined in the repository:*** *In the* ***Created by*** *field, type the user name, which is usually different from the user name. The user name is listed in the principalID attribute on the Attributes subtab of the browse page for the user.* ***To limit the search based on the date the objects were created:****Select an option from the* ***Created on*** *list.****To limit the search based on the user who last modified the objects:****In the* ***Modified by*** *field, type the user name.* ***To limit the search based on the date that the objects were last modified:****Select an option from the* ***Modified on*** *list.*

Click **Search**. A list of search results is displayed on the Search Results page.

In the list, click an object name to view the object, its relationships, and its attributes



Custom attributes for categories and terms

Administrators can define additional custom attributes for categories and terms.

Administrators can create custom attributes to store information about terms and categories, when that information does not fit into the standard attributes and relationships of the glossary model. You can use custom attributes to apply governance standards, enable architecture frameworks, or provide other metadata that is standard for your organization.

When you create a custom attribute, you specify that it applies to either terms or categories, or to both terms and categories. If you apply the custom attribute to both terms and categories, two separate custom attributes are created, one that applies to terms, and one that applies to categories.

Each custom attribute has a name, a description, and a valid value type. The valid value type can be any string or an enumerated list of string values.

You can change the valid value type for a custom attribute at any time. When you change the type, the change does not affect any values that are currently assigned for the attribute. The change determines what will happen the next time a user edits the value for a custom attribute. If you change the type of a custom attribute to String, when users subsequently edit the attribute for any object, they can enter any string value. If you change the type of a custom attribute to Enumerated, when users subsequently edit the attribute for any object, they must select values from the enumerated list of values.

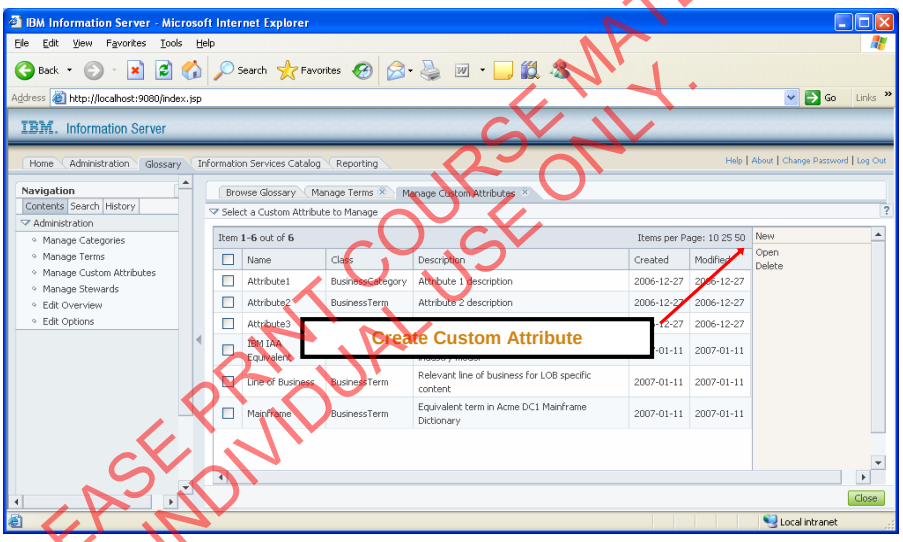
The value of the custom attribute for any particular term or category is initially null. After you create the custom attribute, you can specify its value separately for each term or category that it applies to.

For example, you might create a custom attribute named Data Sensitivity with the following description

A number from 1 to 5, which indicates the sensitivity of the data. Sensitivity is a subjective measure of the impact of the data being released to unauthorized consumers.

You can specify that Data Sensitivity attribute applies only to terms. You choose the enumerated valid value type and enter the numbers 1 through 5 as valid values. After you create the custom attribute, you choose one of those valid values for each particular term that you want to specify a value for.

Creating Custom Attributes



Stewardship

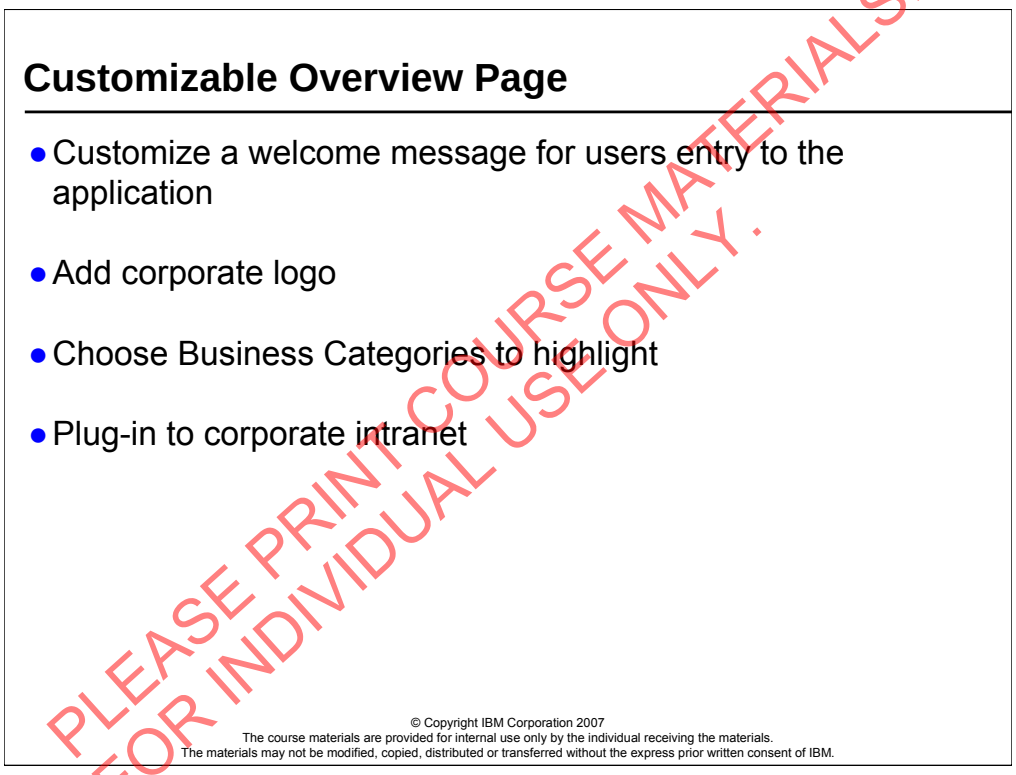
- Determine, document, and share responsible party information for all information integration assets
- Set stewardship for any object
  - Categories
  - Terms
  - Other technical artifacts

Stewards are users or groups that have responsibility for one or more metadata objects in the repository. Business Glossary administrators can designate that a user or group in the metadata repository is a steward. Administrators and authors can then specify that the steward is responsible for one or more metadata objects. A steward is typically assigned to the objects that the user or group is responsible for managing or is the appropriate contact for. When you view the browse page for an object that has a steward, a link to the steward is displayed. The link leads to contact information, which includes e-mail address and phone number. You can assign responsibility for multiple objects when you designate a new steward or when you edit a steward on the Manage Stewards page. You can also assign an object to a steward from the Tasks list on the browse page of the object, or on the browse page of a user or group who is a steward. In addition, you can assign responsibility for a particular category or term to a steward when you create or edit the category or term. An object can have only one steward

Data Steward Role Descriptions

Technology Business Facing Roles	Responsibilities
Lead Data Steward (Data Governance Council)	Add here
Data Steward	Manages the logical data resource or entity of one or more subject areas for the business. Coordinates data definitions, aliases, quality controls, improvement efforts, access authorization, and planning for the subject area data. Establishes data quality metrics. Provides metadata content about the subject area data.
Systems Analyst	Has experience in and is knowledgeable about the relevant IT and User systems used in the business and the processes required to extract the data from these systems. Understands how the data is processed, flows through the systems, and is used by the business from a Technology perspective.

Roles of the Data Steward will vary by organization requirements and resources.



Editing the Overview page

Administrators can edit the Overview page to display a welcome message, company logo, and a list of important categories.

Prerequisite: You must have the Business Glossary Administrator role to perform this task. The header and welcome text that you enter on the Overview page are displayed with line returns and white space intact. You do not need to code special characters for HTML. For example, to display the text Sales & Marketing, you do not need to write Sales & Marketing.

To edit the Overview page:

In the Navigation pane on the Glossary tab, choose **Contents** > **Administration** > **Edit Overview**.

On the Edit Overview page, edit the text and graphics fields that are displayed on the Overview page:

- Type the Overview page headline.
- Type welcome text that includes the corporate message or glossary description to be presented on the page.
- Type the URL to an image to display. The URL must point to an image on the Web accessible to users of the IBM® Information Server Web console. The default image size is 100 by 100 pixels. Larger images will be scaled to this size.

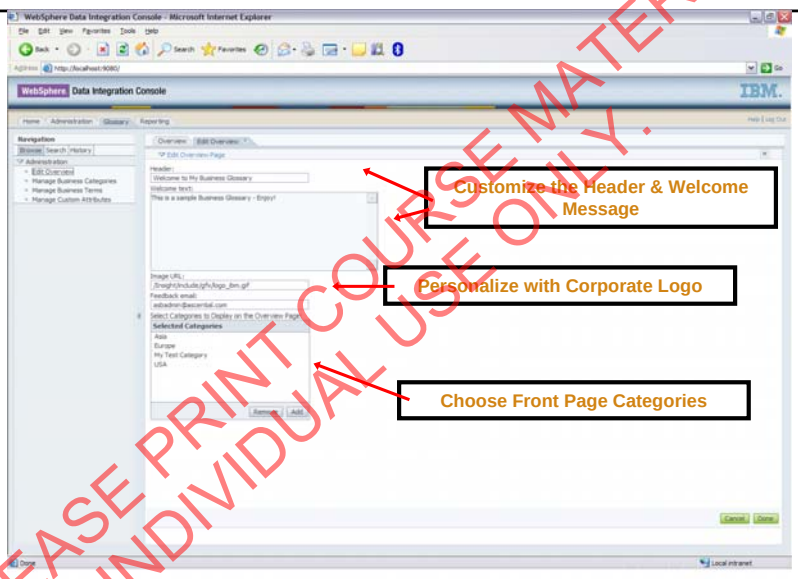
- Type the e-mail address that users should send feedback to. This is presented to users as a "mailto;" hyperlink to the address you specify, based on the default e-mail program of the computer they use. Ensure that all your Business Glossary users can access this address.

Specify the categories to display on the Overview page:

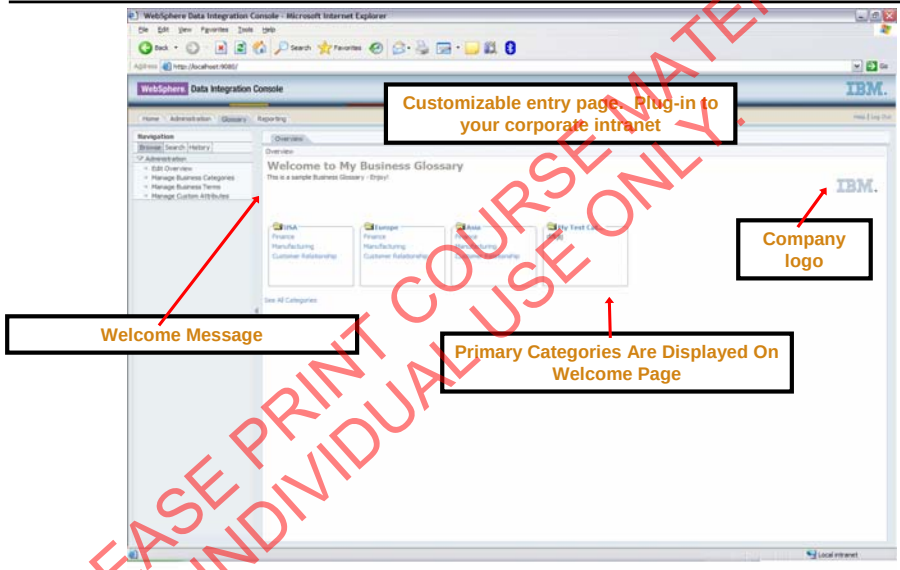
- To add categories to the display, in the **Selected Categories** list box, click **Add** to open the Add Categories window.
- To remove categories from the display, in the **Selected Categories** list box, select one or more categories, and click **Remove**.

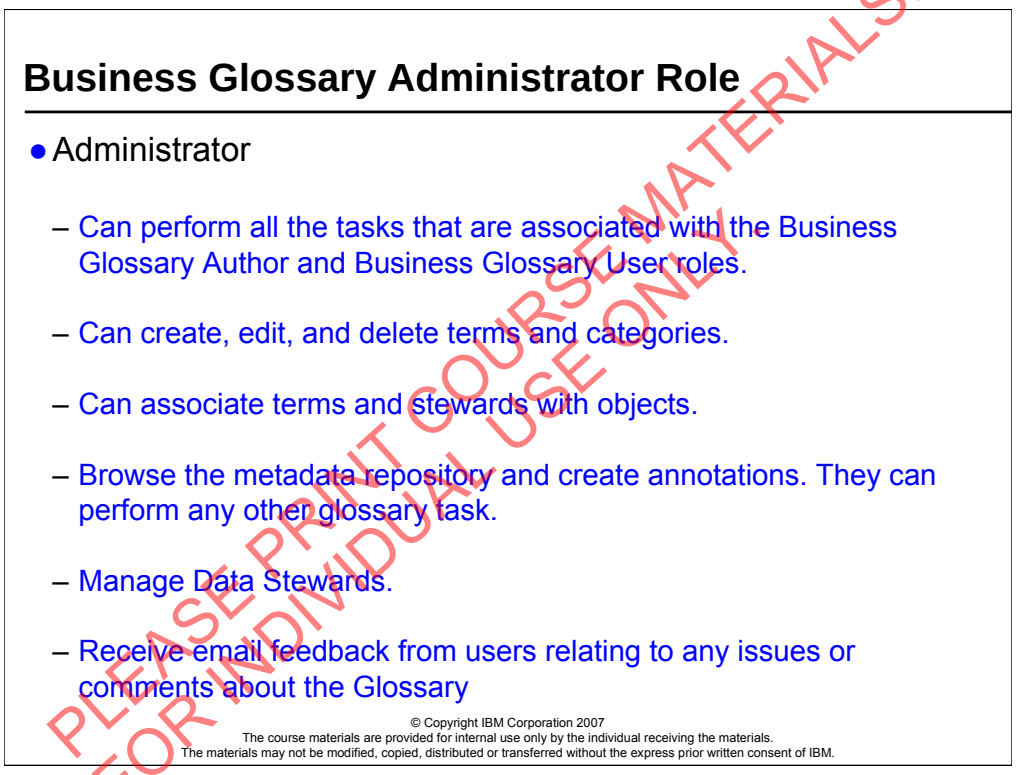
Click **Save and Close**.

Customizing the Overview Page

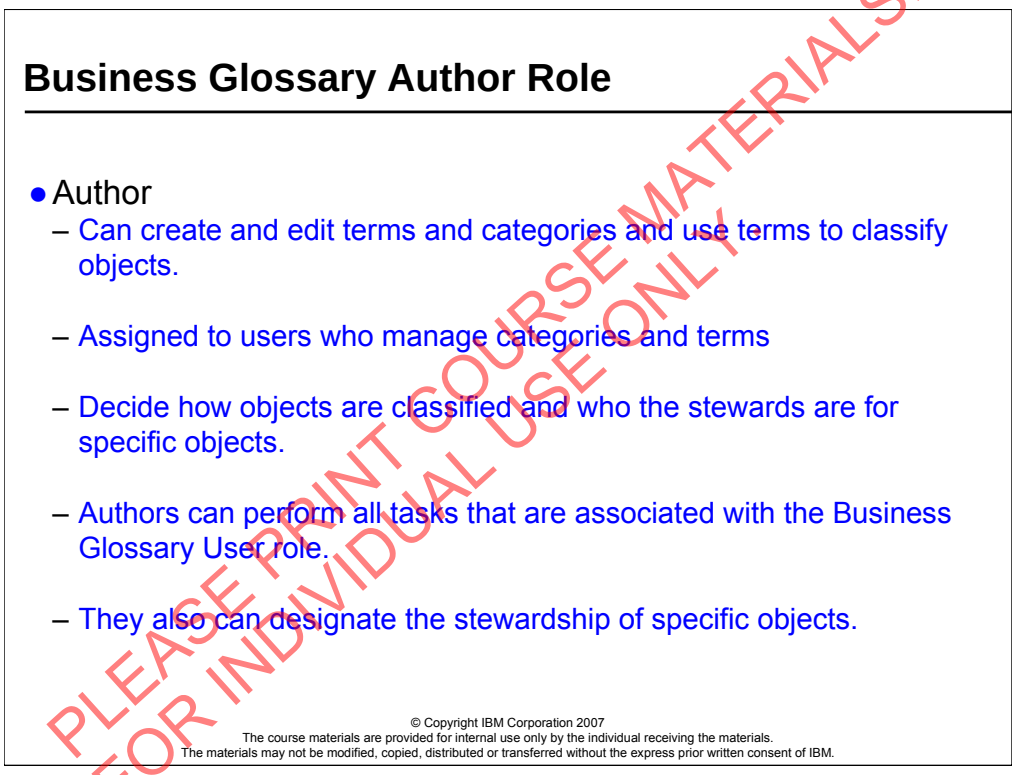


Overview Page

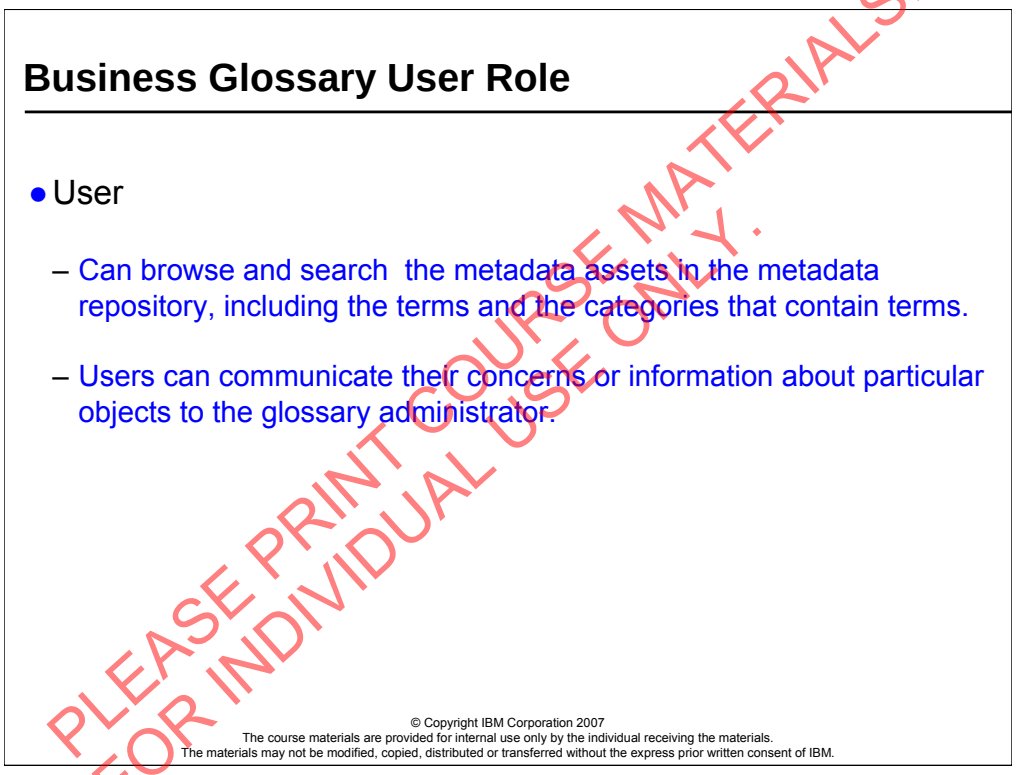




Business Glossary Administrator role Users who are assigned the Business Glossary Administrator role can set up and administer the glossary so that other users can find and analyze the information they need. Glossary administrators can perform all the tasks that are associated with the Business Glossary Author and Business Glossary User roles. They can create, edit, and delete terms and categories. They can associate terms and stewards with objects. They can browse the metadata repository and create annotations. They can perform any other glossary task. In addition, the following tasks can be performed only by people who are assigned the Business Glossary Administrator role: Customizing the Overview page of IBM Information Server Business Glossary to provide users with an starting point that is specific to your enterprise, and that lets them easily navigate the hierarchy of categories. Setting application options. Designating users and groups as stewards, and deleting the steward relationship from a user or group. Creating, editing, and deleting custom attributes. Editing and deleting annotations that were created by others. Deleting terms and categories that were created by others. **Related tasks** “Classifying objects”. Administrators and authors can specify that a term classifies one or more objects in the metadata repository. You classify objects when you are creating or editing terms. “Uploading categories, terms, and custom attributes” Administrators and authors can import files of categories, terms, and custom attributes into the metadata repository. Authors can create and edit categories that contain or reference terms and that serve as subcategories or parent categories to other categories. Administrators and authors can create and edit terms to categorize one or more metadata objects in the metadata repository. Administrators can edit the Overview page to display a welcome message, company logo, and a list of important categories. “Specifying values for custom attributes”. Administrators and authors can specify values for custom attributes for any individual term or category that the attributes apply to. You specify these values when you create or edit the term or category. Administrators and authors can assign responsibility for one or more objects in the metadata repository to stewards.



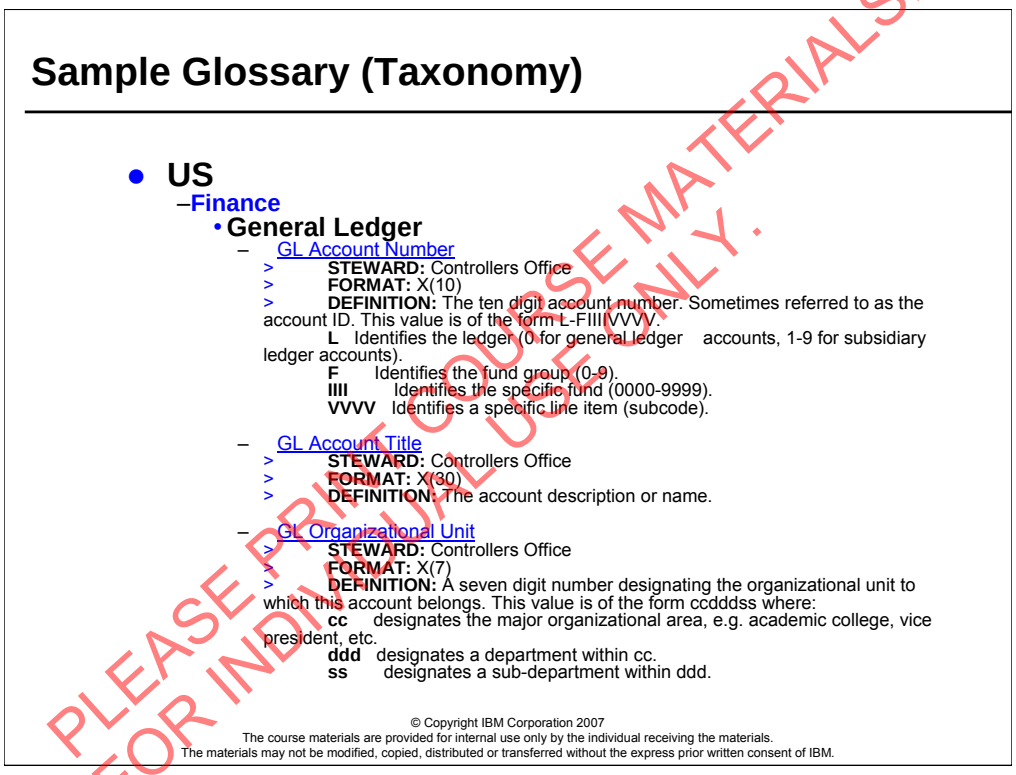
Business Glossary Author role Users who are assigned the Business Glossary Author role can create and edit terms and categories and use terms to classify objects. The author role is assigned to users who manage categories and terms and who decide how objects are classified and who the stewards are for specific objects. Authors can perform all tasks that are associated with the Business Glossary User role. In addition, authors can perform the following types of tasks: Creating and editing a hierarchy of categories that contain terms that are used by your enterprise. Classifying objects in the metadata repository by using terms. Setting stewardship for objects in the metadata repository. Uploading terms and categories to the metadata repository. Specifying values for custom attributes. Administrators and authors can specify that a term classifies one or more objects in the metadata repository. You classify objects when you are creating or editing terms. Administrators and authors can import files of categories, terms, and custom attributes into the metadata repository. Administrators can create custom attributes to store additional information about terms and categories. Administrators and authors can create and edit categories that contain or reference terms and that serve as subcategories or parent categories to other categories. Administrators and authors can create and edit terms to categorize one or more metadata objects in the metadata repository. Administrators can designate any user or group in the metadata repository as a steward. Administrators and authors can specify values for custom attributes for any individual term or category that the attributes apply to. You specify these values when you create or edit the term or category. Administrators can customize settings for IBM Information Server Business Glossary, including the number of objects that are returned by a search. Administrators and authors can assign metadata objects to users and groups that have been designated as stewards. You can assign objects when you create or edit stewards.



Business Glossary User role People with the Business Glossary User role can examine the metadata assets in the metadata repository, including the terms and the categories that contain terms. Users can communicate their concerns or information about particular objects to the glossary administrator. Users can perform the following types of tasks: Browsing the structure of categories and terms. Searching the metadata repository for categories, terms, and other objects. Exploring the attributes and relationships of all objects in the metadata repository. Sending feedback to the administrator. You can use the Overview page to navigate to all top-level categories and their subcategories. You can then explore the categories, their related terms, and other related objects. From the browse page of an object, you can explore the properties and relationships of the object, and read user notes about the object. You can use multiple criteria when you search for objects that are stored in the metadata repository. You can search for objects that are stored in the metadata repository based on the object name or description. From the browse page of an object, you can send feedback about the object to the glossary administrator

Agenda

- Discussion of Business Metadata
  - Introduction and Benefits
  - Metadata Primer
  - Theory & Tools
- Product Overview
  - Purpose
  - Features & Functions
- **Exercises**
  - **Accessing the Glossary**
  - **Administration**
  - **Uploading Categories and Terms**
  - **Browsing and Searching**
  - **Business Category and Term Creation and Editing**
  - **Annotations**
  - **Custom Attributes**



Sample glossary of business terms showing hierarchy and stewardship.