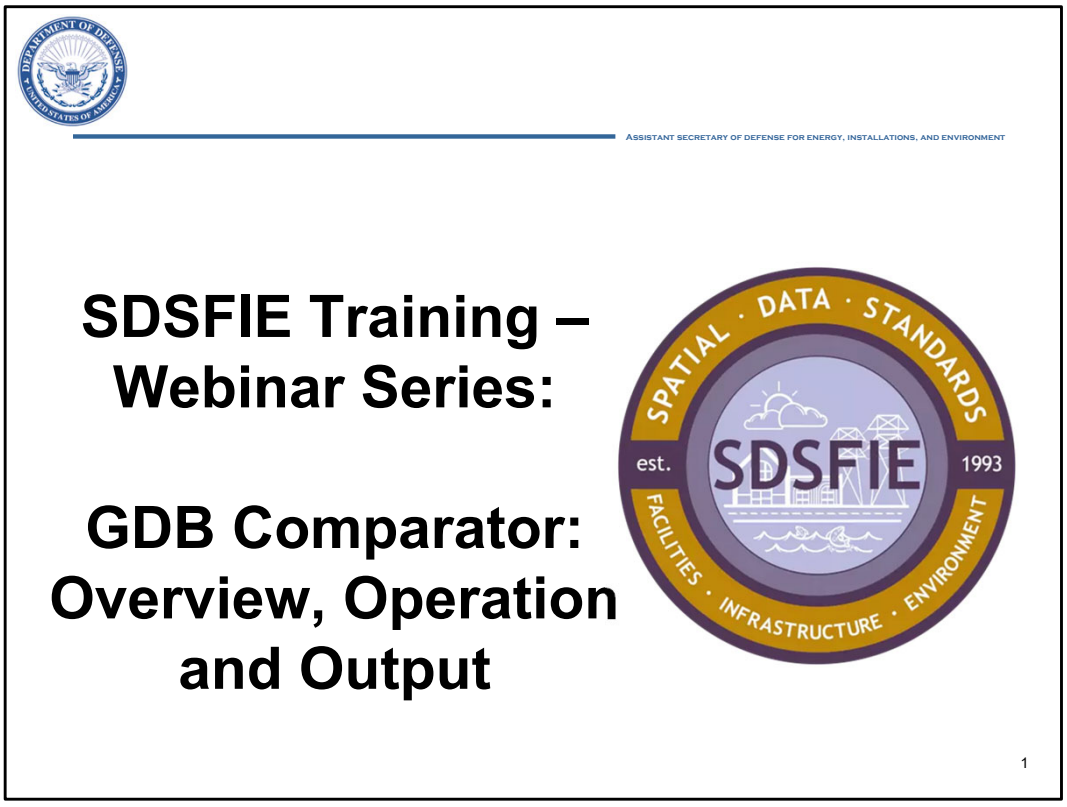
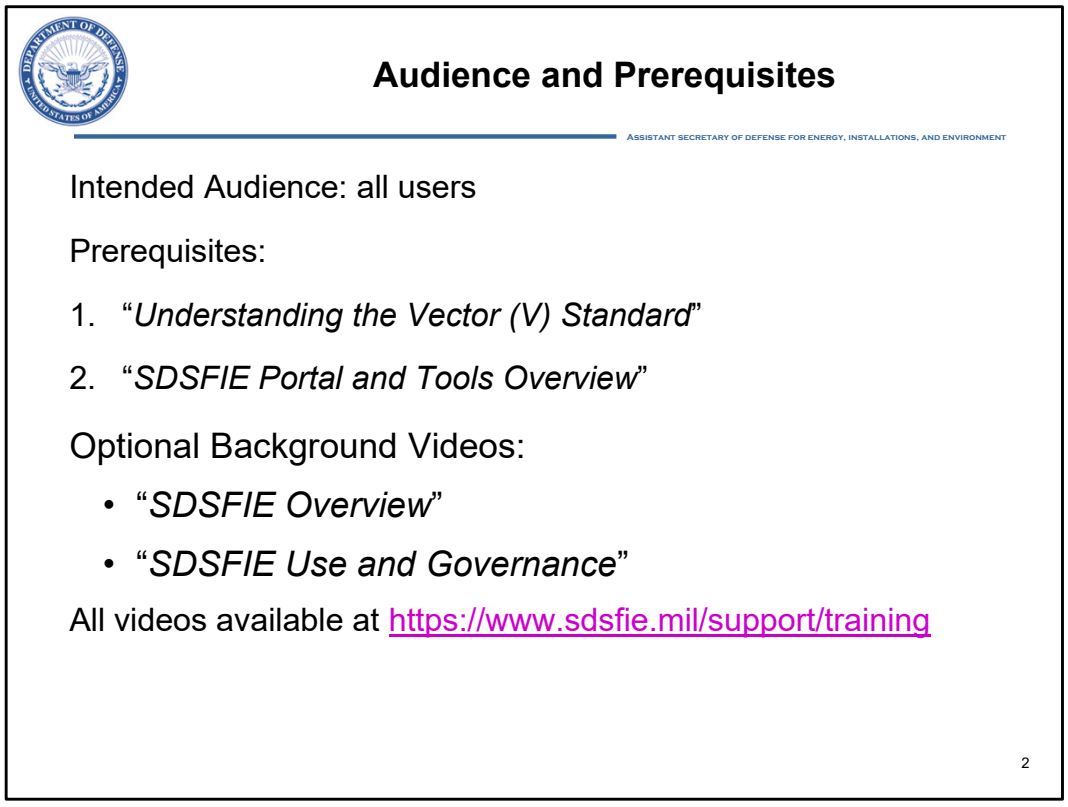



Welcome. This is the “GDB Comparator: Overview, Operation and Output” webinar, a resource for SDSFIE Training.

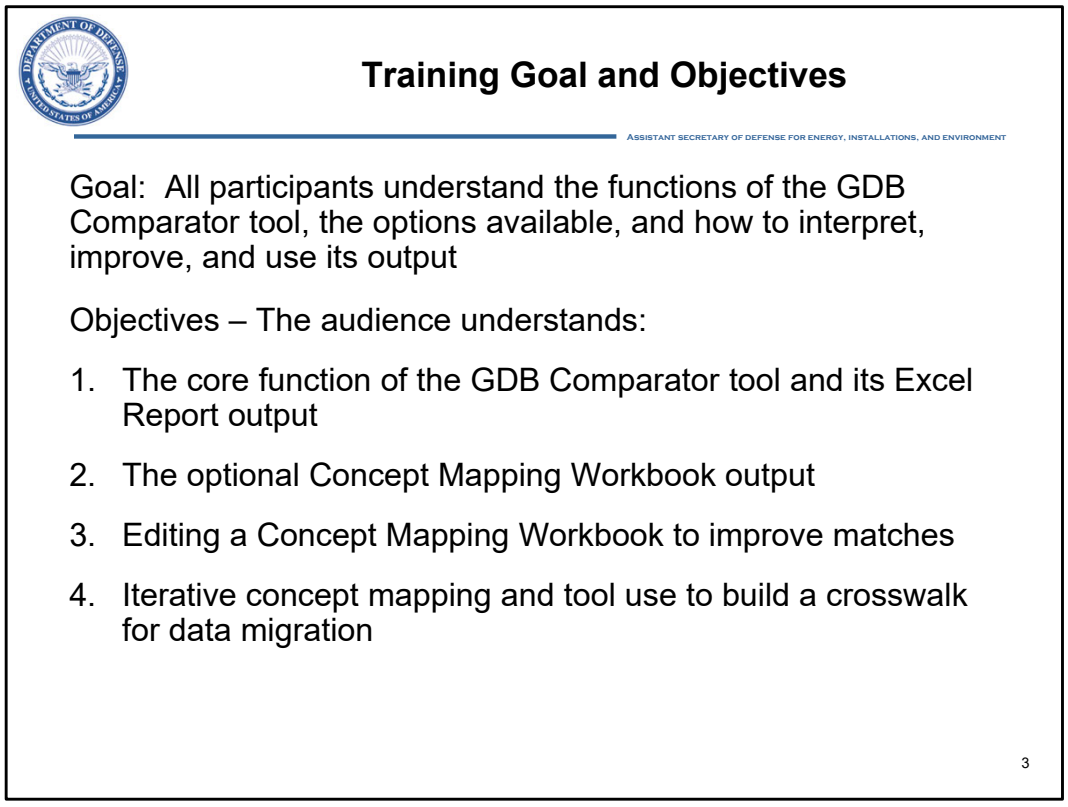



This training resource is designed for all users.

There are two recommended prerequisites for this webinar, watching the "Understanding the Vector Standard" and the "SDSFIE Portal and Tools Overview" video.

The 2 background training resources listed here are optional, but they may add to your overall understanding of SDSFIE.

All of these videos are accessible from the Support section of the SDSFIE Portal at the URL shown.




The overall goal for this training resource is for users to understand the basic and advanced functionality of the ArcGIS "GDB Comparator" tool, and how the output can be interpreted, edited, and used in larger workflows. The acronym G-D-B stands for geodatabase.

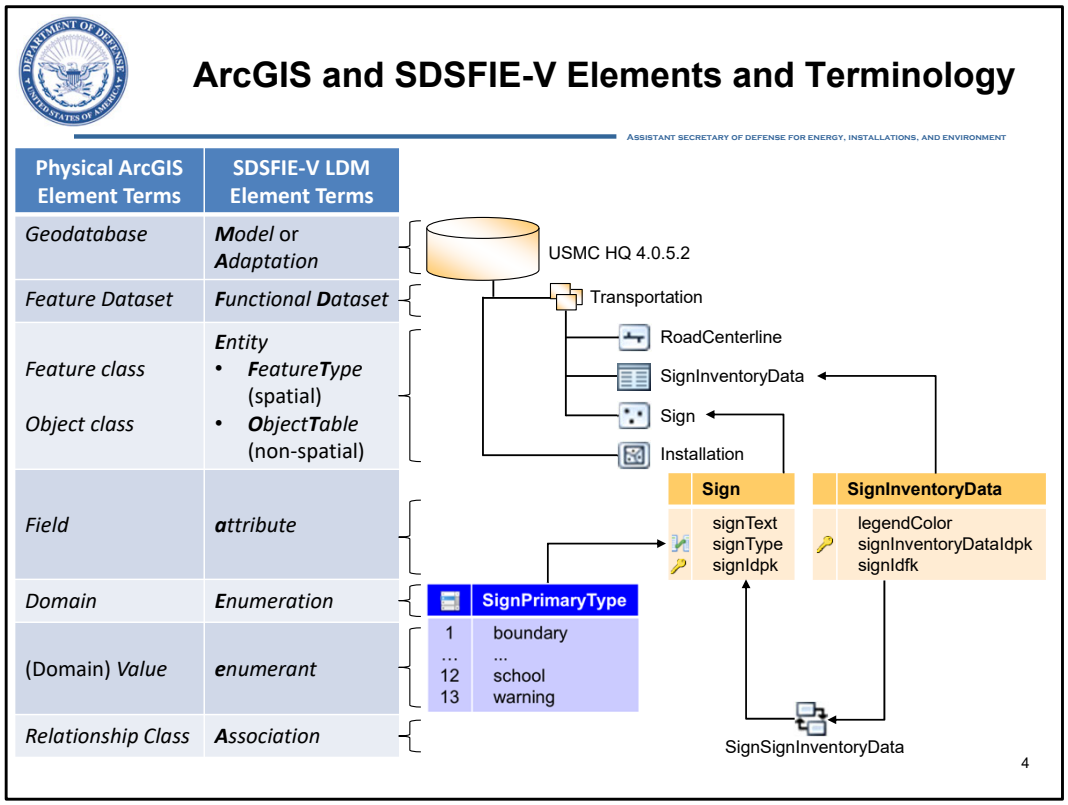
The specific objectives are that the audience gains an understanding of the following:

1. What the GDB Comparator does and the core Excel Report that it produces

2. What the Excel Concept Mapping Workbook output is and what it contains

3. How the Concept Mapping Workbook can be edited to improve the matched elements and then used as an input for running the tool again

4. How iterative Concept Mapping Workbook updates and GDB Comparator tool runs can be used to create an initial crosswalk for data migration in Production Mapping




Using this tool involves extracting elements from ArcGIS geodatabase schemas, abstracting out physical detail such as the geometric specificity of feature classes, and reporting on the elements and their properties in a format much like that in the SDSFIE-Vector logical data model. For this reason, the tool documentation and outputs contain element names and labels from both ArcGIS physical and SDSFIE-V logical lexicons. Given this, let's take a moment to review those corresponding element names down through the element hierarchy.

*click*

- A *model* is simply a specific collection of SDSFIE-V elements. In Esri ArcGIS, models are implemented as geodatabases.

*click*

- A *functional dataset* is a grouping of feature types. In ArcGIS, this corresponds to a feature dataset.

*click*

- In SDSFIE-V there are two types of *entities*: First, *feature type* represents a set of thematically equivalent features which can be expressed as records, or rows in a feature attribute table, that are linked to a geometric feature that is a point, line, or polygon. Second, *object tables* define a non-spatial data table that can be linked to one or more feature type records. In an Esri geodatabase, these elements are called feature classes and object classes.

*click*

- An *attribute* of a entity is a characteristic of the features, and can be expressed as a column in a feature attribute table. In Esri, an attribute is called a field.

*click*

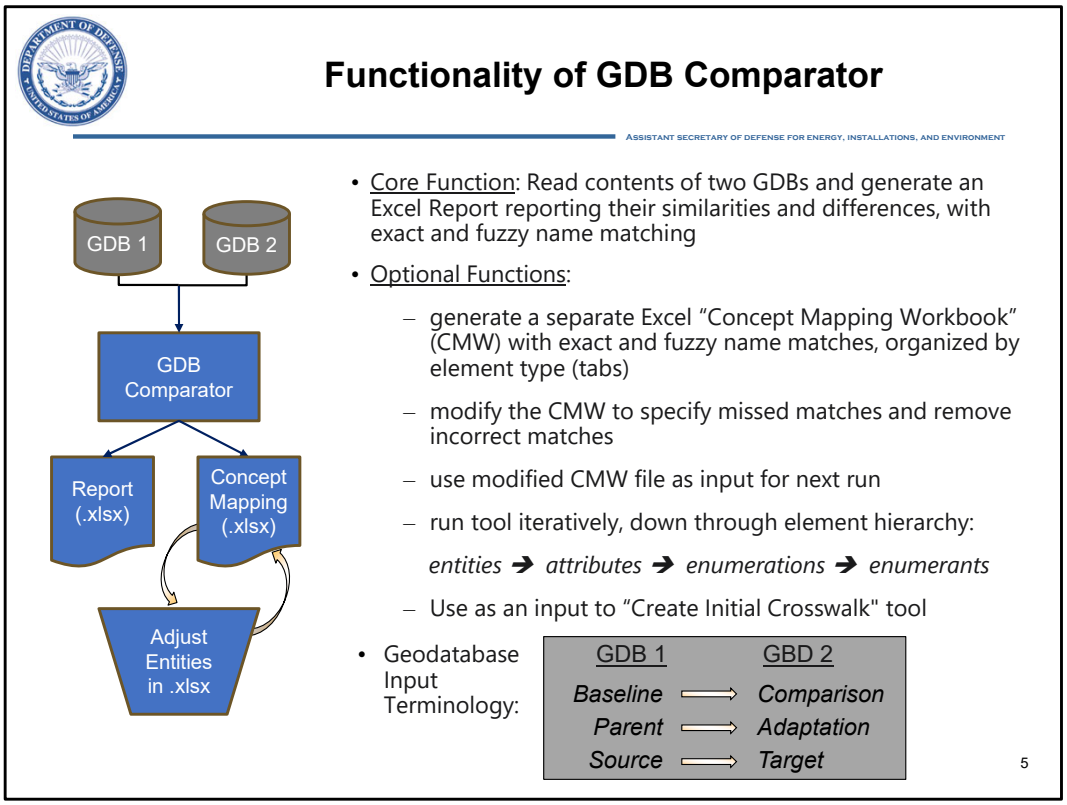
- An *enumeration* constrains an attribute to a permissible set or numerical range of values. In an Esri geodatabase, the synonymous term is domain.

*click*

- An *enumerant* is a value within an enumeration. The enumerants for an enumeration comprise the permissible set of values for that enumeration. In Esri these are called domain values.

*click*

- An *association* is used to establish a relationship between two entities using a foreign key in one attribute table that points at the primary key of another entity. In an Esri geodatabase, this is called a relationship class.




The core functionality of the GDB Comparator tool is to read all of the elements from the schemas of any two geodatabases, make matches between the two sets of element names for each element type, and then generate a report that specifies all elements and where they do or don't match and summarizes the level of alignment by element type. Fuzzy matching algorithms may be used to identify non-exact but likely matches.

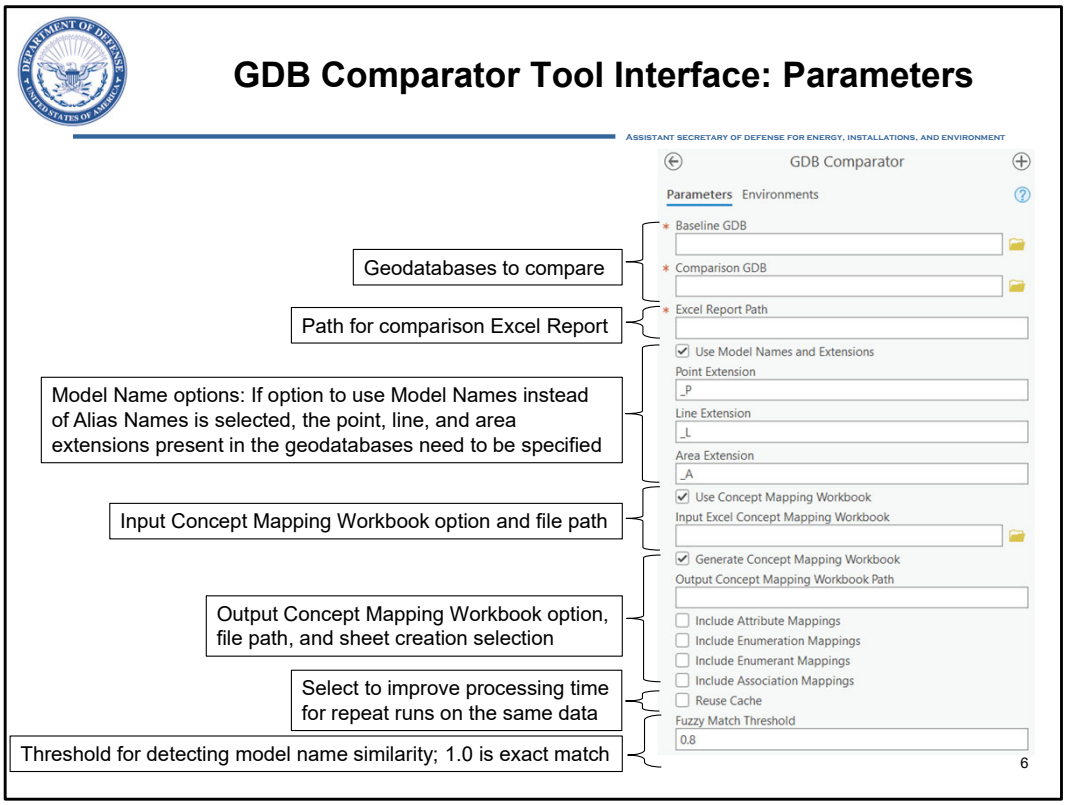
GDB Comparator also has optional functions that revolve around the *Excel Concept Mapping Workbook, or CMW*. *click* This tool output contains much of the same content as the core Excel Report, but allows for editing to improve the fuzzy matching by specifying missed matches and removing incorrect matches, and by specifying entity splits and merges in schema migration.

*click* The edited *CMW* can be used as an input for the tool in an *iterative* manner going down through the model element hierarchy, starting with improving the fidelity of entity matches, then improving on the initial set of attribute matches, and so on until all valid element matches in the model have been established.

At that point, *click* a finished *CMW* can be used as an input to the "Create Initial Crosswalk" tool as part of a data migration workflow.

*click* Regarding the terminology used for its two geodatabase inputs that are generically depicted as "GDB 1" and "GDB 2", while the tool has a general utility of identifying the similarities and differences of any two "Baseline" and "Comparison" geodatabases, its application in the SDSFIE community warrant that these two geodatabases *click* at times be called "Parent" and "Adaptation".

When used in the context of migration workflows, *click* the names "Source" and "Target" may be used.




With an understanding of the tool's core and optional functionality now in hand, and before starting the demo, let's briefly consider how this relates to the tool interface and the parameter value input boxes and check boxes.

Starting at the top, *click* we see that the first 3 input boxes are mandatory, as indicated by the red asterisk: the first GDB, the second GDB, and the filepath and name of the core Excel Report of element matches.

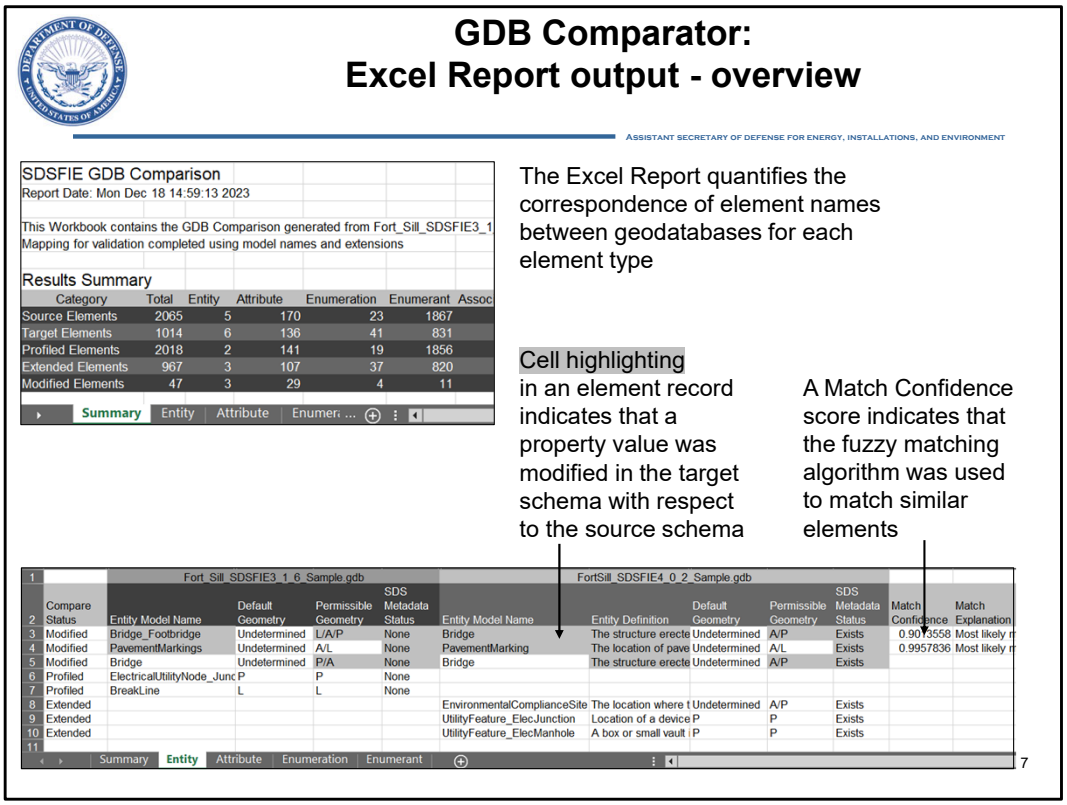
*click* The next section provides a checkbox with a choice to match on element Alias Names or Model Names. If checked, Model Names are used and geometry name extensions are specified in order to conflate to a single feature type name for comparison and mapping.

*click* In the next box the user can specify a Concept Mapping Workbook to be used as an input to improve the subsequent mapping output.

*click* Next, the user can specify the filepath and name of the optional Concept Mapping Workbook output, and checkboxes for which element mappings – each in a separate tab – it should contain.

*click* Below that section, the "Reuse Cache" checkbox offers the user time savings for repeated runs by reusing the cached schema information from each of the two geodatabases.

*click* Finally, the aggressiveness of the fuzzy matching can be specified. A value of 1.0 input here will only make matches if the element names are exactly the same. As shown, the default is set at 0.8, which provides for non-exact matching without too many incorrect matches. We'll now look at some examples using that setting.




Let's also take a quick look at the layouts of and some example records from the Excel-based outputs of the GDB Comparator.

*click*

Excel Reports quantify the correspondence of element names between geodatabases, presented in a separate tab for each element type.

There is also a Summary tab, as shown here in the upper left, with total values of the summary tab of an example Excel Report is displayed.

Looking at the Entity tab from the same example, we see matches for the first 3 records,

*click*

with highlighted or shaded cells where properties differ in some manner between the matched Source and Target feature classes.

Also note that, for name matching that is less than perfect,

*click*

a Match Confidence level between the "1.0" level and the specified Fuzzy Match Threshold is listed. Take the second record for example: the pair of Model Names PavementMarkings and PavementMarking differ only by a single letter, which results in a 0.996 match confidence.

# GDB Comparator: Excel Report output – categorization of record-level correspondence

ASSISTANT SECRETARY OF DEFENSE FOR ENERGY, INSTALLATIONS, AND ENVIRONMENT

SDSFIE GDB Comparison
Report Date: Mon Dec 18 14:59:13 2023

This Workbook contains the GDB Comparison generated from Fort_Sill_SDSFIE3_1
Mapping for validation completed using model names and extensions

Results Summary

| Category | Total | Entity | Attribute | Enumeration | Enumerant | Assoc |
|---|---|---|---|---|---|---|
| Source Elements | 2065 | 5 | 170 | 23 | 1867 | |
| Target Elements | 1014 | 6 | 136 | 41 | 831 | |
| Profiled Elements | 2018 | 2 | 141 | 19 | 1856 | |
| Extended Elements | 967 | 3 | 107 | 37 | 820 | |
| Modified Elements | 47 | 3 | 29 | 4 | 11 | |

Summary | Entity | Attribute | Enumer...

| Category | Source GDB | Target GDB | Match | Identical Properties |
|---|---|---|---|---|
| ***Profiled*** | present | absent | no | n/a |
| ***Extended*** | absent | present | no | n/a |
| ***Modified*** | present | present | yes | no |
| ***Unchanged*** | present | present | yes | yes |

Element record categorization: based on record-level name matching and property differences

Source (Parent) GDB - Entity Presence or Absence

Target (Adaptation) GDB - Entity Presence or Absence

| | Fort_Sill_SDSFIE3_1_6_Sample.gdb | | | | FortSill_SDSFIE4_0_2_Sample.gdb | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Compare Status | Entity Model Name | Default Geometry | Permissible Geometry | SDS Metadata Status | Entity Model Name | Entity Definition | Default Geometry | Permissible Geometry | SDS Metadata Status | Match Confidence | Match Explanation |
| 3 | Modified | Bridge_Footbridge | Undetermined | I/A/P | None | Bridge | The structure erecte | Undetermined | A/P | Exists | 0.9013558 | Most likely n |
| 4 | Modified | PavementMarkings | Undetermined | A/L | None | PavementMarking | The location of pave | Undetermined | A/L | Exists | 0.9957836 | Most likely |
| 6 | Profiled | ElectricalUtilityNode_Junc | P | P | None | | | | | | | |
| 8 | Extended | | | | | EnvironmentalComplianceSite | The location where t | Undetermined | A/P | Exists | | |
| 10 | Extended | | | | | UtilityFeature_ElecManhole | A box or small vault | P | P | Exists | | |

Summary | Entity | Attribute | Enumeration | Enumerant

8

In the Excel Report output, each element is examined from the standpoint of a parent and adaptation model, and categorized according to the 4 modeling decisions available in the adaptation process.

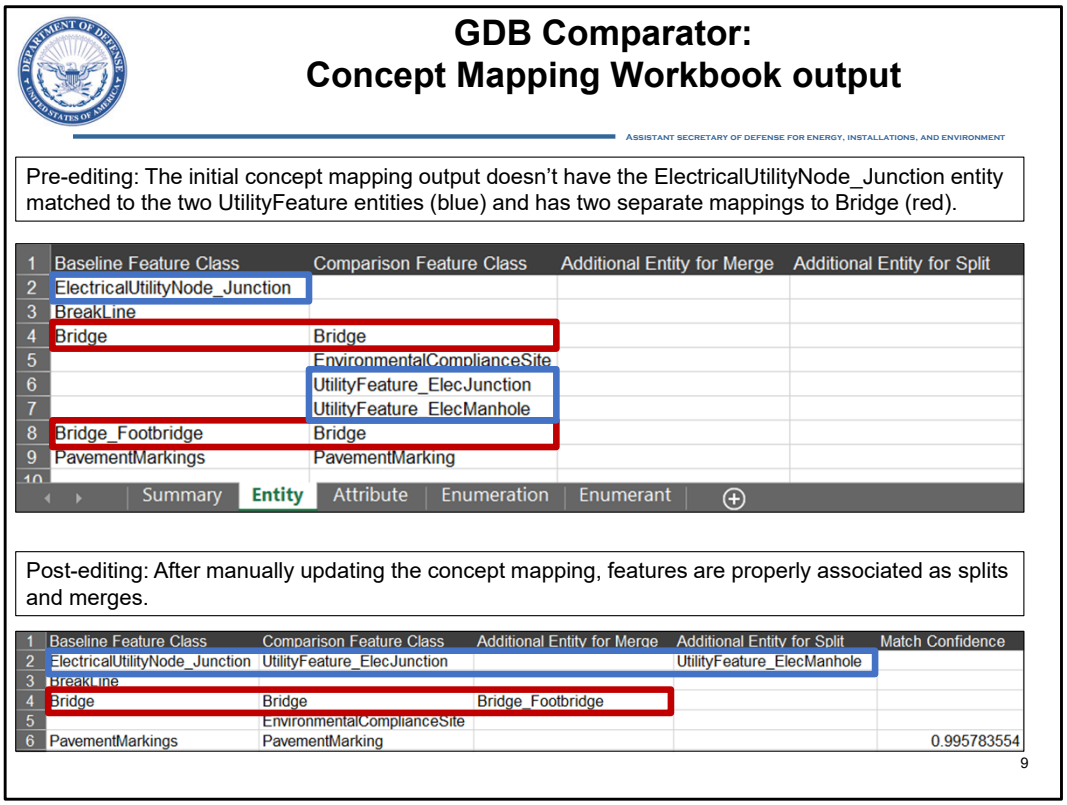
Matched elements, *click* those that are present in both the source (or parent) schema, and target (or adaptation) schema, *click* can be either “Modified” or “Unchanged”. Unmatched elements that are present only in the source *click* *click* are termed to be “Profiled”, whereas unmatched elements present only in the target *click* *click* are termed “Extended”.

In the example shown here, there are Entity records for all of these 4 categories except Unchanged.

The PavementMarking entity match *click* is of the orange Modified type, given the name change from plural to singular.

The ElectricalUtilityNode_Junction entity is of the purple Profiled type, *click* given that it is present in only the parent geodatabase schema.

Conversely, EnvironmentalComplianceSite is only present in the adaptation schema, *click* and is thus of the green Extended variety.

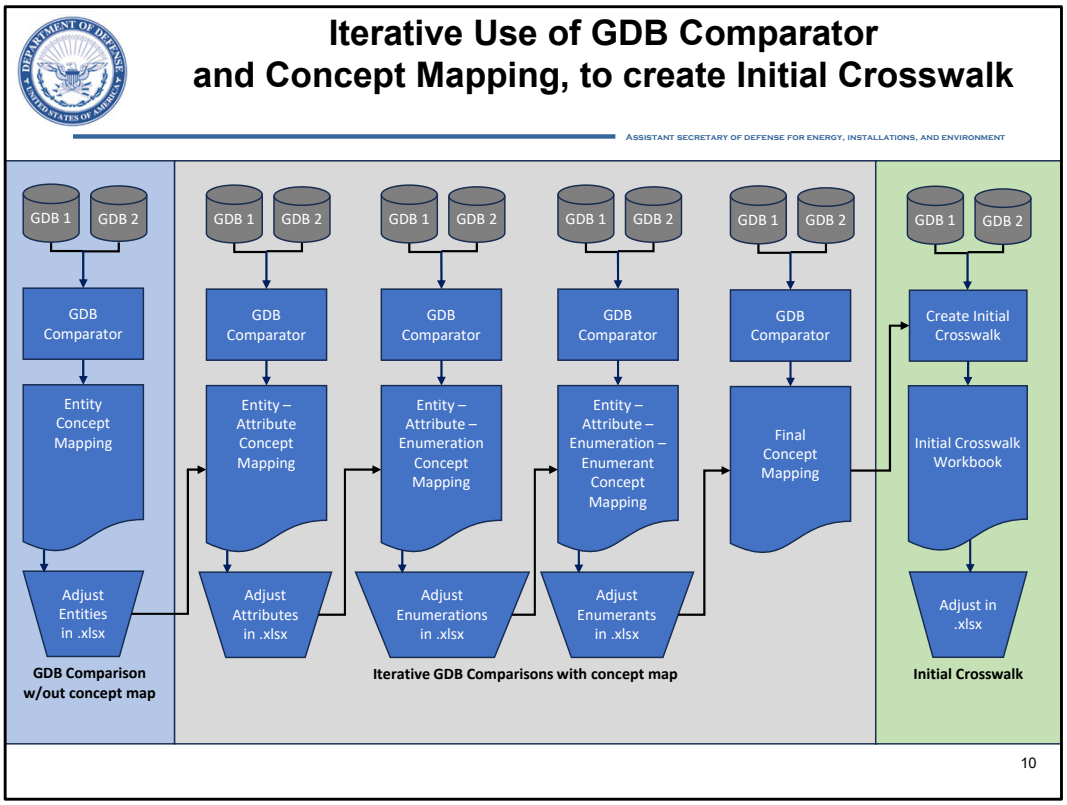



The other type of Excel-based output of the GDB Comparator tool is the Concept Mapping Workbook. The examples here highlight the fuzzy matches present in the workbook prior to editing, and the changes made to improve the match fidelity.

In the pre-edited Entity tab, there are mapping improvements that can be made for two sets of entities.

First, the ElectricalUtilityNode_Junction entity is not matched to the two UtilityFeature entities to which it should be split (as shown in the blue highlights), and second, there are two separate mappings to the Bridge entity in the target which should be represented as a merge.

The edits are made directly in the workbook by moving – but never completely re-moving – entities. The result is the post-editing condition show below, with the single ElectricalUtilityNode_Junction source entity being split into two UtilityFeature_ElecJunction and UtilityFeature_ElecManhole target entities, and two Bridge and Bridge_Footbridge source entities being merged into a single Bridge target entity.




This diagram depicts the use case described previously, in which the GDB Comparator is run iteratively, each subsequent time using an updated Concept Mapping Workbook to adjust matches in the next lower level of the element hierarchy.

Note that the example just discussed on the previous slide, in which matches in the Entity tab were adjusted and splits and merges were established, is illustrated in the blue shaded area at the left hand side of this diagram and at the start of chain.

The updated Concept Mapping Workbook with its adjusted Entity tab contents will serve as an input to the process step to the right, *click* on the far left-hand side of the gray shaded area. In this process step, given that entity matches are correct, the pre-edited fuzzy matches of next lowest level in the element hierarchy – the attributes – can then be evaluated and edited as necessary.

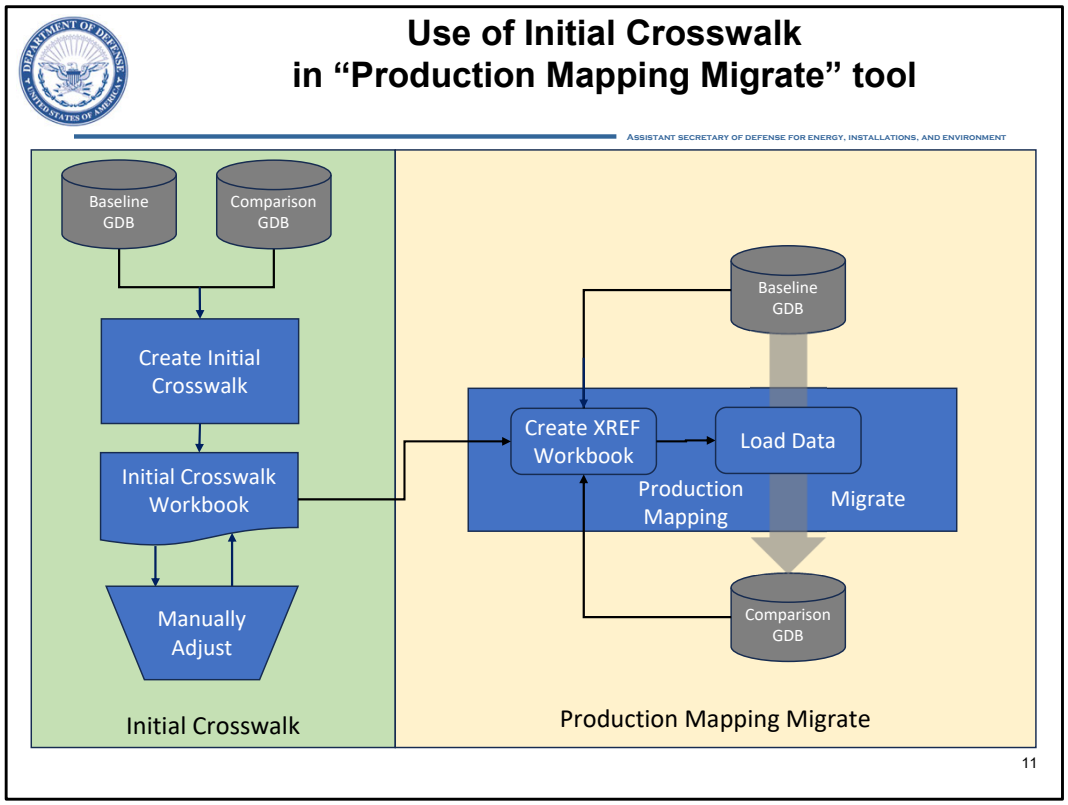
Then that workbook with improved entity and attribute matching serves as an input to the next iteration. This pattern holds for evaluating and editing the enumeration matches in the middle-left process, *click* and the enumerants in the middle-right process, *click* until a final Concept Mapping Workbook is produced from the rightmost process in the gray shaded area.

*click*

At this point, that final workbook can serve as an input *click* to the Initial Crosswalk Workbook tool.

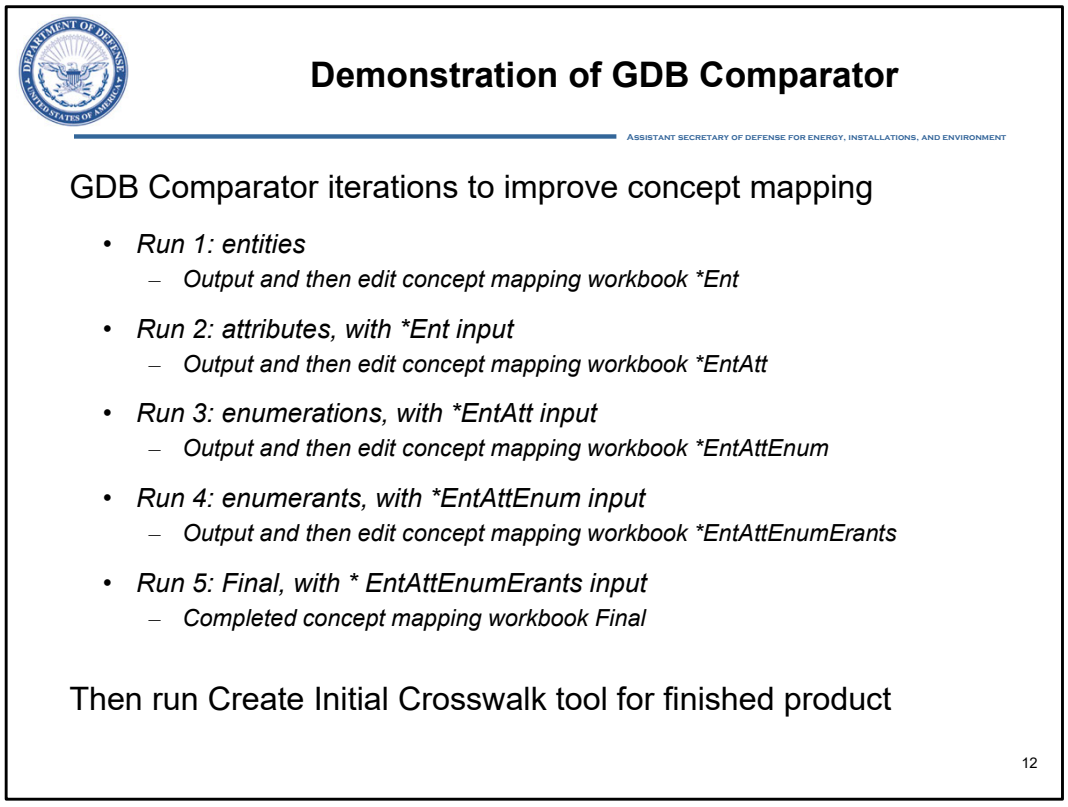
This is the use case that will be demonstrated in a moment.

.




But, to finish the journey, we need to consider the "Production Mapping Migrate" tool. This last tool in the migration workflow chain provides the means to the create the Cross-Reference or "XREF" workbook from the initial Crosswalk Workbook endpoint of the iterative GDB Comparator tool use just illustrated.

The XREF workbook defines that actions taken in the final Load Data step of the Production Mapping workflow, where data records are actually moved from the Baseline or Source GDB to the new schema that is defined in the Comparison or Target GDB.




As depicted previously, 5 GDB Comparator iterations are required to create the input for the Create Initial Crosswalk tool. We will now follow this process, creating matches for each level in the element hierarchy to be examined, validated, edited, and used as an input in the next iteration of the process.

We begin the demo with the same two geodatabases used as examples in the slides: the schemas for a subset of the feature classes at Fort Sill – the source in the Army H-Q 3.1.6 Adaptation and the target in the Army H-Q 4.0.2 Adaptation.

Here in ArcGIS Pro in the Catalog Pane on the right, an expanded view of the two GDBs are presented. Note the underscore-P, underscore-L, and underscore-A geometry-based extensions used for the feature class Model Names. As we will see in a moment, those groups of feature classes will be conflated in the concept mapping output of the GDB Comparator Tool.

The SDSFIE ArcGIS Toolbox has already been downloaded from the SDSFIE Portal, and we can add it to this project now.

This toolbox contains all 3 of the tools that we just described in the diagrams of the migration workflow. We will start the first iteration of the GDB Comparator now.

In the tool interface, let's start filling in the fields.

First, we'll select the Baseline and Comparsion GDB inputs.

Next we will specify the Excel Report filename as "GDBComparison_FortSill1.xls"

The, after checking the "Generate Concept Mapping Workbook" box, we will set the output Component

Mapping Workbook, or CMW, filename to “ConceptMapping_FortSill_Ent.xls”

At this point we will not include mappings for any of the elements other than the entities, which is done by default. We will check the “Include element mappings” boxes in subsequent iterations.

Notice that the “Cataloging feature classes and tables” step takes most of the time. For geodatabases with larger and larger numbers of entities, this is increasingly true. Next time we will check the “Reuse Cache” option to skip that step by reusing the cache that was just created.

This is the Summary tab of the Excel Report that the tool produced. We see here that although conflating the geometry-based Model Name extensions has reduced the number of entities down to 5 in both the Source and Target GDBs, there are a large number of other elements further down in the hierarchy – attributes, enumerations, and enumerants – to be matched.

Let’s take a quick look at the Entity tab before moving directly to our concept mapping iterations.

If we hide some of the columns, we see that 3 entity-level Model Name matches have been made:

- One exact match in row 3 for the Bridge entity [highlight the row], and two fuzzy matches in rows 4 and 5

We will explore these same results further, where we can edit them, in the Concept Mapping Workbook

Here in the Entity tab of the Concept Mapping Workbook, or CMW, we see the same 3 Model Name matches that were present in the Excel Report:

- One exact match in row 3 for the Bridge entity
- Two fuzzy matches in rows 8 and 9.

Both the Bridge and Bridge_Footbridge entities in the source match to the single Bridge entity in the target.

This can be interpreted as a “merge” action, so we will move the Bridge_Footbridge entity to the “Additional Entity for Merge” column in row 3, and delete row 9.

Conversely, only ElectricalUtilityNode_Junction is present in the source, but both UtilityFeature_ElecJunction and UtilityFeature_ElecManhole are present in the target. We will interpret this as a split, and move UtilityFeature_ElecJunction up to row 4 to make it the match, and move UtilityFeature_ElecManhole to the “Additional Entity for Split” column in row 4, and delete rows 6 and 7

Now we will save the file, adding an “underscore-Edited” suffix so we can differentiate it from its unedited precursor.

It is now ready for use as an input to the next iteration of the tool. With entities properly matched, we can move on to the attributes.

Referring to the workflow diagram, we have completed the left-most actions and are ready to use the edited CMW as an input to the 2nd iteration of the GDB Comparator tool.

For iteration 2, we can just edit the parameters used previously. We will leave the Excel Report file name unchanged. First, check the “Use Concept Mapping Workbook” box, and then navigate to the Excel file

that we just saved and select it.

Note that you may need to press the refresh button to see the Excel files.

Second, and as indicated in the pre-demo briefing, we will name the CMW output from the second iteration with an E-n-t-A-t-t suffix, and then check the “Include Attribute Mappings” box.

Finally, we will check the “Reuse Cache” box to speed up the process, as mentioned.

Running the script, notice the speed at which the output is produced by skipping the feature class cataloging steps.

As evident here in the Entity tab, the changes to the entity mappings that we made in the last version of the CMW were preserved. And as we specified, the CMW now includes an Attribute tab where we have 2 tasks for every set of attributes for matched entities: one, validate the fuzzy matches made and correct where necessary, and two, add new matches where necessary.

The source Breakline entity was unmatched and thus “profiled”, and all of the Breakline attributes are profiled as well.

Next up is Bridge, where there are several matches to check. Sorting the Bridge by the “Match Explanation” field, we can easily review just the matches.

A reasonable assumption is that all of the “Exact word matches” are valid, and we can focus our attention on the fuzzy matches.

All four of these matches are for foundational attributes, and all 4 are valid, with the differences simply due to Model Name conventions changing from the 3.0 to 4.0 SDSFIE Gold models. So now, with all of the matches validated, we can look to see if there are any additional matches to be made.

There are 27 unmatched attributes on the source side, and only 8 on the target side, so an efficient approach would be to examine the unmatched target attributes.

If we knew about the lineage of the Bridge attribution in the Army HQ model, we would know that these matches could be established:

- the source bridgeType matches the target nbiDesign
- bridgetypeNBI matches nbiServiceType
- predominantDesignUse matches rpaPredomCurrentUseCatCode, and
- interestType matches rpInterest

So we move these 4 source attribute Model Names down to their matching target attribute and delete the unmatched row that they were in.

Now, we have to do this for the attributes of the other matched entities, and the PavementMarkings match is next.

As we did for the Bridge match, we will sort the PavementMarkings to PavementMarking attribute rows by the “Match Explanation” field and look at the fuzzy matches first.

Here again we see 4 valid matches for foundational attributes.

But there is an error:  the source rpuid matched the target rpsuid at a 90.9% Match Confidence. This is a good lesson to look carefully at the fuzzy matches and know that the 0.8 setting can be overly aggressive in some circumstances, such as this pair of attributes with Model Names that differ only in 1 character.

If we look just above this error, in row 92, we see that the source rpsuid was also matched to the target rpsuid, of course correctly in that case.

The solution is to delete the target rpsuid in row 95, along with matching information to the right.

Next, we would look at the unmatched attributes, for which there are 6 from the source and 3 from the target.

Here again, expertise with the lineage of these adaptations is necessary to make further matches. In this case we know that the source markingColor attribute can be matched to the pavementMarkingColor attribute.

Next, we would follow the same process for the sets of attributes belonging to the other 3 matched entities. For the sake of brevity, we won't show those details.

Once this task is complete, we save the file, again adding an "edited" suffix.

It is now ready for use in the next iteration of the tool. With entities and attributes now properly matched, we can move on to the enumerations and then enumerants.

We have now completed iterations 1 and 2, and are ready to use the edited CMW as an input to the 3rd iteration of the GDB Comparator tool.

For iteration 3, we will edit the parameters used previously. First, we need to change the input Concept Mapping Workbook file to the one with the entity AND attribute information, that we just saved.

Again, you will likely need to press the refresh button to see the new Excel files.

Second, we will name the CMW output from the third iteration with an Ent-Att-ENUM suffix, and then check the Include Enumeration Mappings box and run this third iteration of the tool.

Examining the Attribute tab of the output CMW, and scrolling down to the Bridge to Bridge matches, we can see that the matches we established were preserved.

Moving to the Enumerations tab, we see that there are only 4 matches made: 3 exact and 1 fuzzy.

Given that the manual matching processes are largely the same for the next two element types, and in the interest of time, we will move on to the state at which we have validated and updated all of the matches for all of the element types in our two geodatabases.

With iteration 3 of the GDB Comparator tool and the Enumerations now complete, we will use the edited CMW as an input to iteration 4.

There, after checking the exact and fuzzy Enumerant matches, and establishing new ones where necessary, the last edited CMW file is ready for use for iteration 5.

This last iteration of the GDB Comparator tool outputs a final CMW with correctly mapped Entities, Attributes, Enumerations, and Enumerants. That file is itself a stand-alone crosswalk between the elements of the two Geodatabases.

If, however, we want to utilize the migration workflow supported with SDSFIE Portal tools we will use that final CMW as an input for the Initial Crosswalk Workbook tool.

From that tool interface, we simply:

- Name the crosswalk GDBComparatorDemo\SDSFIE_Initial_Crosswalk_for_FortSill.xls
- Check the Use Concept Mapping Workbook box
- Select the "FINAL" CMW GDBComparatorDemo\Concept_Mapping_FortSill_FINAL.xls as that CMW input, and
- Then press Run.

Examining the Entity tab of the Initial Crosswalk Workbook output, we see that it contains the "FROM" and "TO" feature class names from out source and target geodatabases, respectively. All of our matches are applied as specified with the Migrate, Merge, and Split "Migration Action" assigned. Our unmatched source entity is marked as "Delete" and our unmatched target entity is marked as "Add New".

This sheet and the others in this workbook are aligned to the form of the Production Mapping XREF Excel workbook or to a format from which the XREF sheets can be populated in an automated manner.

Looking at the Attribute tab for the PavementMarkings to PavementMarking crosswalk, we see that our matches, and our single unmatch of the source rpuid and the target rpsuid, were all maintained. The actions here, given that we did not assign any attribute splits or merges, are only Delete, Migrate, or Add New.

Note there is no "Enumeration" tab. This is because, in an XREF file, all of the information required for migrating enumerant values is contained in a single sheet.

Note that in this sheet there is information about the "FROM" or source, and "TO" or target, for all of the elements in the hierarchy: entity, attribute, enumeration, and enumerant. In the actual XREF sheet, all of this information is simplified to numeric codes for the matched entities and their matched attributes that have constraining enumerations, along with the possible source and target Enumerant values.

We just produced an Initial Crosswalk Workbook. That file, in turn, serves as an input to the Production Mapping Migrate tool, and creates an XREF Workbook which provides all of the mappings required to load data from the populated Source GDB to the empty Target GDB using ArcGIS Production Mapping processes.

You now know how to approach creating, in a semi-automated manner, element-level mappings between two geodatabases which can then define the migration actions in a fully-automated data loading procedure.

This concludes the demonstration portion of the video.


This concludes the "Geodatabase Comparator: Overview, Operation and Output" training.

If you want to learn more about SDSFIE, its Governance and Use, or more detailed SDSFIE topics, visit our Training Resources page at the URL listed here.

Thank you for your interest in SDSFIE.