

DATA RECOGNITION ACCURACY

IBM GUARDIUM DISCOVER AND CLASSIFY

VERSION 4.2.2

TABLE OF CONTENTS

Table of Contents	2
Overview	3
Master Catalog	3
Data Quality Evaluation	3
Catalog Re-creation	3
Chapter 1: Data Quality Evaluation	4
Data Quality in Personal Information Search	4
Chapter 2: Data Discovery and Classification Accuracy Report	11
Overview	11
Prerequisites for Data Elements validation	12
Prerequisites for Classifier validation	13
Data validation	14
Accuracy report export	16
Chapter 3: Catalog Recreation	20
Catalog Recalculation	21
Recalculation History	23

OVERVIEW

Data quality in the IGDC catalog of personal data depends on multiple factors like data quality in the sources of trusted data, data recognition configuration, constraints for recognition of unique data subjects in your network, etc.

This guide describes how to evaluate the accuracy of data discovery in various IGDC modules, identify sources with low data quality, and recalculate the master catalog after fixing the data recognition configuration.

MASTER CATALOG

IGDC analyzes the data at rest, data in motion and reference data (root data assets) within your network and aggregates all the discovered personal data in a master catalog with the following information:

1. Personal data records grouped by unique persons. Each record is an aggregation of personal data elements grouped by category. There are two types of values in the record:

- **Trusted** - root data assets and values validated against root data assets;
- **Discovered** - values unconfirmed against root data assets.
- **Conflict** - unique data element with 2 or more values discovered. A data element is marked as unique in the Data element configurator, meaning that the system expects this data element to have only one value. For example, a person does not have two SSN numbers or birthdays in real life. The following unique data elements are defaulted: birthday, driver license, national ID, passport number, and tax ID.

In addition to the discovered personal data, the master catalog also contains metadata: how many data subject profile copies are in the network, the sources where each data subject has been discovered, including information on the source (Type, URL, etc.), and which data assets the data subject belongs to (business context).

DATA QUALITY EVALUATION

After generation of the master catalog of personal data, IGDC offers the data quality evaluation in the Personal Information Search: users can review the trusted, discovered, and conflict values per specific data subject.

CATALOG RE-CREATION

IGDC offers two options for the master catalog re-creation:

- **Recreate**: Recreates the catalog after data quality improvements, like deletion of RDA, re-upload of a more relevant RDA upload, or update of the Data element configuration.
- **Recreate & Drop**: Recreates the catalog with trusted data only, meaning that the values that were not validated against RDA are removed from the catalog.



Only users with SUPER_ADMIN permissions have access to the **Re-calculation** page. For other roles (admin, user, DSR operator), the **Re-calculation** option is not available in the sidebar menu.



IGDC is designed for operation with the latest version of Google Chrome. Using other browsers is not recommended and may affect performance and functionality.

CHAPTER 1: DATA QUALITY EVALUATION

This chapter describes how to evaluate data quality using the **Personal Information Search** tool.

To have data for quality evaluation, ensure the following :

1. Analyze at least one root data asset.
2. Analyze a data source with data from the applied root data asset.



For information on how to create and manage root data assets, see the **Root Data Asset Management User Guide** and **Supervised AI User Guide**.

For information on how to analyze data sources, see the **Data Source Catalog User Guide**.

DATA QUALITY IN PERSONAL INFORMATION SEARCH



For detailed information on how to use the personal information search tool, see the **Personal Information Search User Guide**.

1. Go to **IGDCPlatform > Data Privacy > Personal Information Search**. In the **Data element (1)** and **Value (2)** fields, configure the search query. For example, **city = maitland**. Then click **Search (3)**.

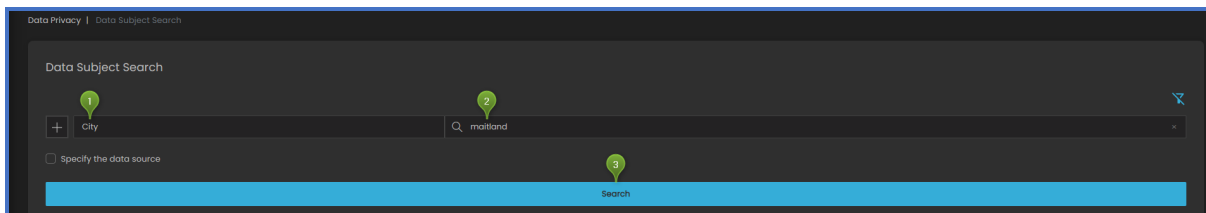


Figure 1: Personal information search landing page

2. On the search results page, click the desired data subject's name to open their profile.

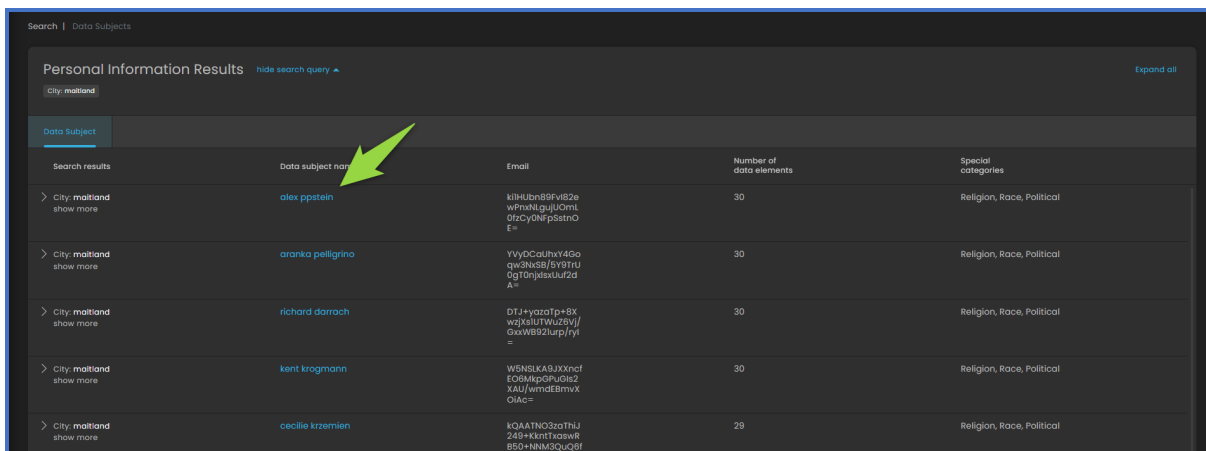


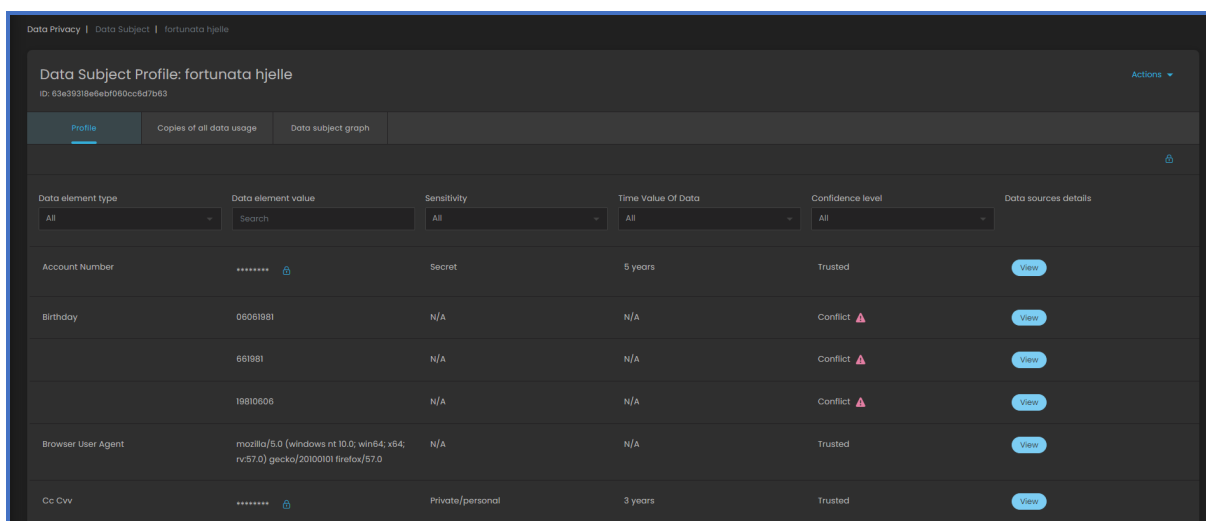
Figure 2: Search results page

3. On the data subject profile page, review the data subject's discovered values in the **Profile** tab. This tab allows you to evaluate the level of trust for each data element available in the master catalog for the data subject: trusted, discovered or conflict value.

Table 3: Properties in the Profile details tab

PROPERTY	DESCRIPTION
Data element type	Type of discovered data. For example, city, given name, credit card number, etc. You

PROPERTY	DESCRIPTION						
	can select the data elements to show using in the Data element dropdown list.						
Data element value	Discovered value. Several values can be discovered for the same data element. You can find the desired value using the Search box.  Note: If the data element is encrypted, the value will be masked (*****) or hashed, with the ability to unlock and view the value by clicking the  (Locked) icon. This user action will be logged.						
Sensitivity	Sensitivity class assigned to the data element. You can configure sensitivity classes on the Sensitivity page (Business Context > Sensitivity) and then assign them to the data element in the Data element configurator (CM UI > Settings > Data Recognition > Data Elements Configuration).						
Time value of data	Time value of the data element configured in the Data element configurator (CM UI > Settings > Data Recognition > Data Elements Configuration).						
Confidence level	Level of trust for the discovered value. You can sort the displayed data elements by confidence level in the dropdown list. <table> <tr> <td>Trusted</td><td>Values discovered in the root data asset - approved source of truth.</td></tr> <tr> <td>Discovered</td><td>Values discovered automatically by IGDC during network or data-at-rest analysis.</td></tr> <tr> <td>Conflict</td><td> Unique data element with 2 or more values discovered. A data element is marked as unique in the Data element configurator, meaning that the system expects this data element to have only one value. For example, a person does not have two SSN numbers or birthdays in real life. Default unique data elements: Birthday, Driver license, National ID, Passport number, Tax ID. To configure uniqueness for the data elements, go to Settings > Data Recognition > Data Element Configuration. For more information, see the Analytic Engine and Console Manager Administrator Guide. </td></tr> </table>	Trusted	Values discovered in the root data asset - approved source of truth.	Discovered	Values discovered automatically by IGDC during network or data-at-rest analysis.	Conflict	Unique data element with 2 or more values discovered. A data element is marked as unique in the Data element configurator, meaning that the system expects this data element to have only one value. For example, a person does not have two SSN numbers or birthdays in real life. Default unique data elements: Birthday, Driver license, National ID, Passport number, Tax ID. To configure uniqueness for the data elements, go to Settings > Data Recognition > Data Element Configuration. For more information, see the Analytic Engine and Console Manager Administrator Guide.
Trusted	Values discovered in the root data asset - approved source of truth.						
Discovered	Values discovered automatically by IGDC during network or data-at-rest analysis.						
Conflict	Unique data element with 2 or more values discovered. A data element is marked as unique in the Data element configurator, meaning that the system expects this data element to have only one value. For example, a person does not have two SSN numbers or birthdays in real life. Default unique data elements: Birthday, Driver license, National ID, Passport number, Tax ID. To configure uniqueness for the data elements, go to Settings > Data Recognition > Data Element Configuration. For more information, see the Analytic Engine and Console Manager Administrator Guide.						
Data sources details	The View button opens a popup with details about the data source. The popup content depends on the value's confidence level.						



Data element type	Data element value	Sensitivity	Time Value Of Data	Confidence level	Data sources details
Account Number	***** 	Secret	5 years	Trusted	View
Birthday	06061981	N/A	N/A	Conflict 	View
	661981	N/A	N/A	Conflict 	View
	19810606	N/A	N/A	Conflict 	View
Browser User Agent	mozilla/5.0 (windows nt 10.0; win64; x64; rv:57.0) gecko/20100101 firefox/57.0	N/A	N/A	Trusted	View
Cc-Cvv	***** 	Private/personal	3 years	Trusted	View

Figure 4: Data Subject Profile page

4. Pay special attention to data elements with multiple values of various confidence levels. They can assist in improving the quality of trusted data sources and master catalog.

In the example below, 4 values have been discovered for the National ID data element:

- **2 conflict values** were retrieved from the trusted data sources - root data assets. As National ID is configured as a unique data element, these values are marked as conflict. This means that one of the root data assets contains incorrect data or National ID is not a unique value in your network.
- **2 discovered values** were identified automatically by IGDC.

CC CVV	933	Trusted	view
Government issued ids			
Driver License	o576565354604	Trusted	view
National id	222254000	Discovered	view
	111254000	Conflict	view
	300254000	Conflict	view
	333254000	Discovered	view
Passport Number	391604026	Trusted	view

Figure 5: Example of a data element with multiple values of various confidence level in Personal information search

5. Click **View** next to the trusted values to explore the page with auxiliary information on the sources of trusted data.

Table 6: Auxiliary data on trusted sources

PARAMETER	DESCRIPTION
ID	Unique identifier automatically assigned to the root data asset by IGDC.
Data element type	Type of discovered data. For example, city, given name, credit card number, etc.
Data element value	Value retrieved from the root data asset. Several values can be discovered for the same data element.  Note: If the data element is encrypted, the value will be masked (*****) or hashed, with the ability to unlock and view the value by clicking the  (Locked) icon. This user action will be logged.
Trusted	Number of root data assets containing the subject data element and value. In the example below, IGDC discovered the city = stamford value for this data subject in the root data asset.
RDA name	Name of the root data asset entered in the Root Data Asset Manager or in Supervised AI.
RDA URL	URL of the root data asset.
Data Source Type	Data source type: file/database.
Schema	Corresponding DB schema.
Table	Corresponding DB table.
Column	Corresponding DB table column.

Data Privacy | Trusted sources for data subject pr... | 653bd8e90e717b2740b80134

Trusted sources for data subject profile

ID: 653bd8e90e717b2740b80134
Data element type: Birthday
Data element value: 071986
Trusted: 2

RD A Name	RD A URL	Data Source Type	Schema	Table	Column
> aqa_admin_ms_sql	jdbc:sqlserver://10.192.192.36:1433	Database	N/A	N/A	N/A
> C11753 RDA	jdbc:sqlserver://10.192.192.36:1433;	Database	AQA_FLAT_IT_I0K_SHORT.dbo	AQA_FLAT_IT_I0K_SHORT.dbo.SEP_VALID_CANDIDAT E	VALID_BIRTHDAY

Figure 7: Trusted sources for Data subject profile page in data subject search

6. Click **View** next to the discovered values to explore the page with auxiliary information on automatically discovered data.

Table 8: Auxiliary data on discovered sources

PARAMETER	DESCRIPTION
ID	Unique identifier automatically assigned to the root data asset by IGDC.
Data element type	Type of discovered data. For example, city, given name, credit card number, etc.
Data element value	Value automatically discovered by IGDC during data source analysis. Several values can be discovered for the same data element.  Note: If the data element is encrypted, the value will be masked (*****) or hashed, with the ability to unlock and view the value by clicking the  (Locked) icon. This user action will be logged.
Discovered	Number of data sources containing the subject data element and value. In the example below, IGDC discovered the National ID = 222254000 value for this data subject in data source.
Data source name	Name of the data source. For example, file name.
Data source type	Data source type: file/database.
Discovered source URL	URL of the data source.
Schema	Corresponding DB schema.
Table	Corresponding DB table.
Column	Corresponding DB table column.

Discovered sources for data subject profile ID: 630dd400ed93536170aeb50c

Data element type: National Id
Data element value: 222254000
Discovered: 1

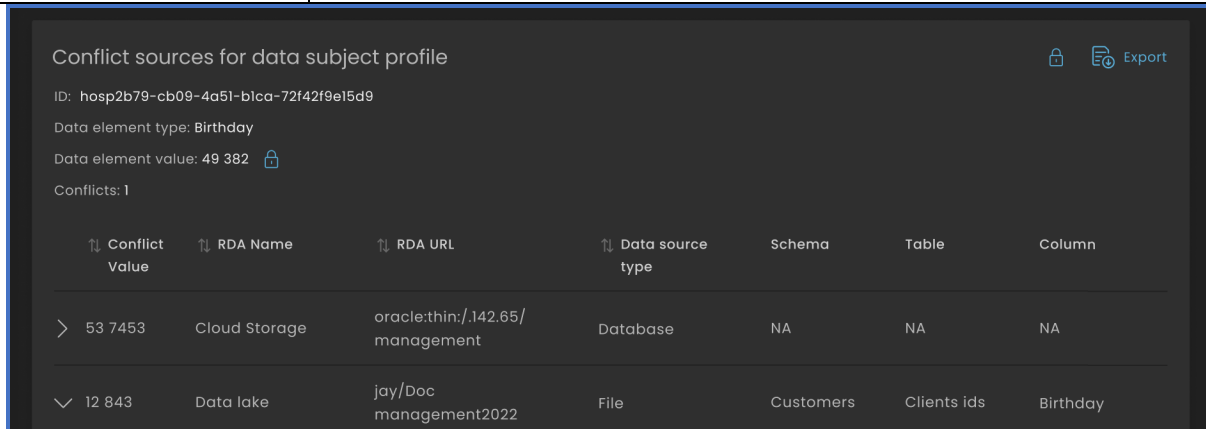
Discovered source name	Discovered source URL	Discovered source type
ct5b	nfs://10.192.192.29//samba/ltouch_share/CO NFLUCT4//ct5b.xlsx	File

Figure 9: Discovered sources for data subject profile in data subject search

7. Click **View** next to the conflict values to explore the page with axillary information on potential root data assets with conflict data and compare the conflicted values.

Table 10: Auxiliary data on conflict sources

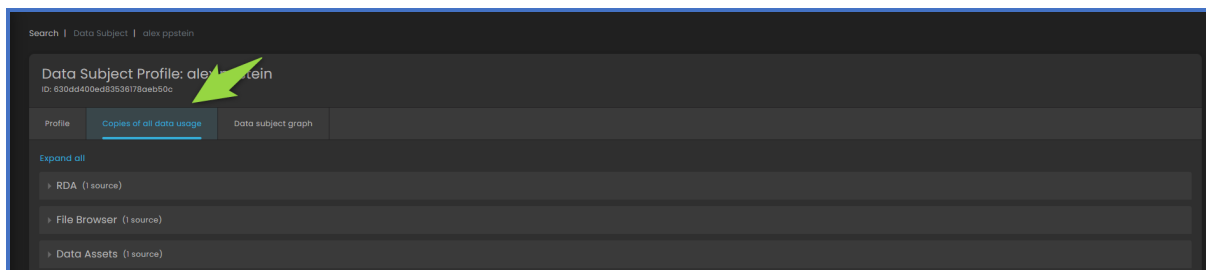
PARAMETER	DESCRIPTION
ID	Unique identifier automatically assigned to the root data asset by IGDC.
Data element type	Type of discovered data that must be unique in the data subject profile. The unique data elements are defined in the Data element configurator (Settings > Data Recognition > Data Element Configuration). Default unique data elements: Birthday, Driver license, National ID, Passport number, Tax ID.
Data element value	One of several values retrieved from this root data asset for a unique data element.  Note: If the data element is encrypted, the value will be masked (*****) or hashed, with the ability to unlock and view the value by clicking the  (Locked) icon. This user action will be logged.
Conflicts	Number of conflict sources containing the alternative value.
Conflict value	Alternative value retrieved from another root data asset for the unique data element.
RDA Name	Name of the root data asset entered in the Root Data Asset Manager or in Supervised AI.
RDA URL	URL of the root data asset.
Data source type	Data source type: file/database.
Schema	Corresponding DB schema.
Table	Corresponding DB table.
Column	Corresponding DB table column.



Conflict Value	RDA Name	RDA URL	Data source type	Schema	Table	Column
> 53 7453	Cloud Storage	oracle:thin://142.65/management	Database	NA	NA	NA
✓ 12 843	Data lake	jay/Doc management2022	File	Customers	Clients ids	Birthday

Figure 11: Conflict sources for Data subject profile in data subject search

8. Go to **Copies of all data usage** tab to see the root data assets and the data sources associated with the data subject. For each data usage type, you can see the number of sources discovered by IGDC.



Data Source	Count
RDA	1 source
File Browser	1 source
Data Assets	1 source

Figure 12: Copies of all data usage tab in the Personal information search tool

9. Expand the **RDA** section and explore the list of root data assets.

Table 13: RDA file parameters in the Copies of all data usage tab in the Personal information search tool

PARAMETER	DESCRIPTION
Date (1)	Date and time the file was updated, e.g., Mar 12, 2018, 12:53:32
Name (2)	Name and format of the document that contains the data subject's personal information, e.g., customers-europe.csv
File size (3)	Size of the file, e.g., 23.43 MB
Special (4)	Number of special categories of the data subject's personal information, e.g., 4 Special Categories

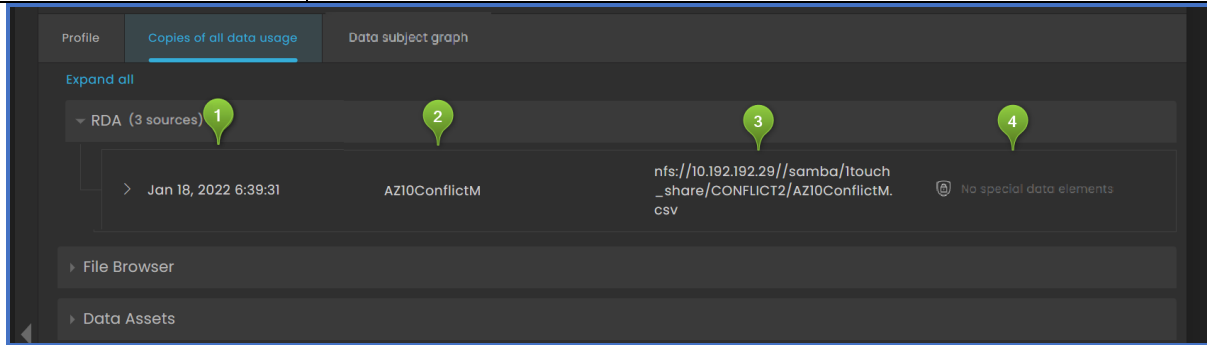


Figure 14: RDA section for files in the Copies of all data usage tab in the Personal information search tool

Table 15: RDA file parameters in the Copies of all data usage tab in the Personal information search tool

PARAMETER	DESCRIPTION
Date (1)	Name of the database that contains the data subject's personal information. e.g., CRM
Tables (2)	Number of the database tables that contain the data subject's personal information.
URL (3)	Database URL
Special (4)	Number of special data elements for the data subject's personal information, e.g., 4 Special Categories

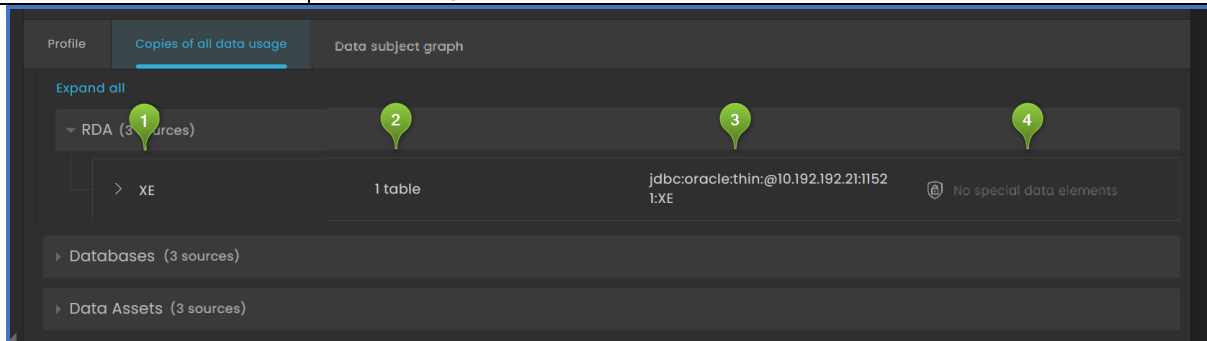


Figure 16: RDA section for databases in the Copies of all data usage tab in the Personal information search tool

12. Expand the RDA file/database to see the data subject's data elements grouped by category extracted from the RDA, as well auxiliary information on the file (name, type, path, size) or database (path, table names).

RDAs (1 Source)

1

2

3

4

5

6

jdbc:mysql://aqa-data-ass...

1 table

jdbc:mysql://aqa-data-ass...

No special data elements

db path: jdbc:mysql://aqa-data-asset-test-stand.ttouch.corp:3306

Data element type	Data element value	Sensitivity	Time Value Of Data	Confidence level	Schema name	Table name	Column name
Birthday	123087	N/A	N/A	Trusted	SISTEM_SMOKE_1 T_10K_FULL	SISTEM_SMOKE_1 T_10K_FULLCAN DIDATE	BIRTHDAY
Given Name	muse	N/A	N/A	Trusted	SISTEM_SMOKE_1 T_10K_FULL	SISTEM_SMOKE_1 T_10K_FULLCAN DIDATE	GIVEN_NAME
Surname	beaubien	N/A	N/A	Trusted	SISTEM_SMOKE_1 T_10K_FULL	SISTEM_SMOKE_1 T_10K_FULLCAN DIDATE	SURNAME

Figure 17: Expanded RDA section for databases in the Copies of all data usage tab in the Personal information search tool

CHAPTER 2: DATA DISCOVERY AND CLASSIFICATION ACCURACY REPORT

OVERVIEW

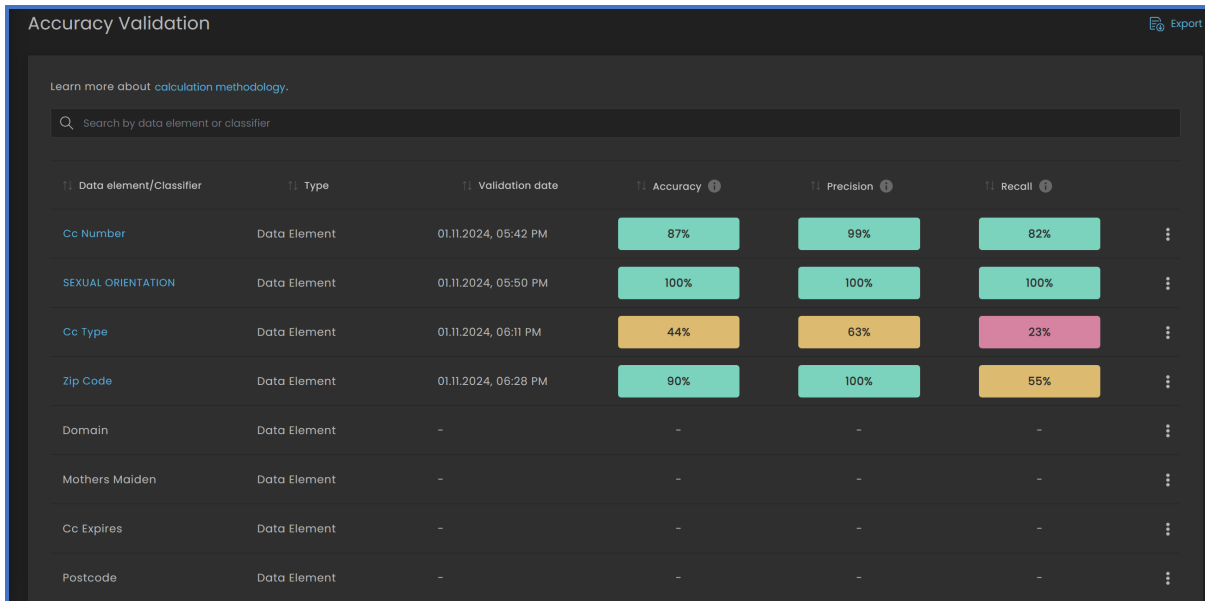
This chapter explains how to generate an export the data accuracy report (XLSX or CSV) that shows how precisely the system recognizes data elements and classifiers. Data accuracy evaluation is conducted based on your provided set of correct and incorrect values.

The data accuracy report is used to evaluate and demonstrate how accurately IGDC is identifying and protecting your organization's data. You can use this report, for instance, during an audit as proof of credit card information protection reliability based on your data set that includes values that are credit card numbers and values that are not (see [Prerequisites for data elements validation](#) for details).

The starting **Data accuracy** page (**Supervised AI > Accuracy Validation**), displays the data elements and classifiers registered in the system along with their last accuracy validation results if there are any.

Table 1: Properties on the data accuracy page

PROPERTY	DESCRIPTION
Data element /Classifier	Title of the data element or classifier.
Type	Type of the data element or classifier.
Validation date	The date and time of the last successful validation.
Accuracy	Percentage of correctly classified instances out of the total instances. Accuracy rate: (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative). In the example above, this measures how well IGDC identifies credit card numbers.
Precision	Percentage of positive predictions that are actually correct. Precision = True Positive / (True Positive + False Positive). Example: If IGDC correctly identifies 80 credit card numbers (True Positive) and incorrectly identifies 20 non-credit card numbers as credit card numbers (False Positive), the precision can be calculated as: Precision = $80 / (80 + 20) = 0.8$ or 80% This means that out of all the instances classified as credit card numbers, 80% are actually credit card numbers, while the remaining 20% are false positives.
Recall	Represents the model's ability to find all the positive instances. Recall rate: = True Positive / (True Positive + False Negative) Example: Out of 100 credit card numbers, IGDC correctly identifies 85 (TP) but misses 15 numbers and classifies them as non-credit card numbers (FN). The recall can then be calculated as: Recall = $85 / (85 + 15) = 0.85$ or 85% This means that IGDC captured 85% of the actual credit card numbers.



Accuracy Validation Export

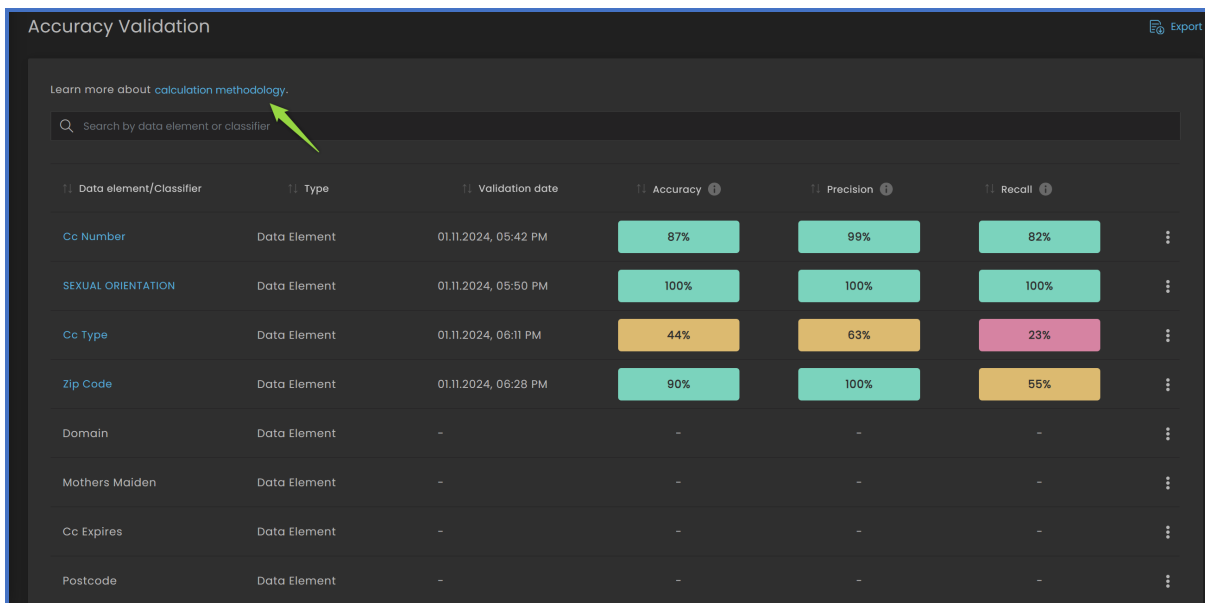
Learn more about [calculation methodology](#).

Search by data element or classifier

Data element/Classifier	Type	Validation date	Accuracy	Precision	Recall	
Cc Number	Data Element	01.11.2024, 05:42 PM	87%	99%	82%	⋮
SEXUAL ORIENTATION	Data Element	01.11.2024, 05:50 PM	100%	100%	100%	⋮
Cc Type	Data Element	01.11.2024, 06:11 PM	44%	63%	23%	⋮
Zip Code	Data Element	01.11.2024, 06:28 PM	90%	100%	55%	⋮
Domain	Data Element	-	-	-	-	⋮
Mothers Maiden	Data Element	-	-	-	-	⋮
Cc Expires	Data Element	-	-	-	-	⋮
Postcode	Data Element	-	-	-	-	⋮

Figure 2: Data accuracy page

Click the **Learn more about calculation methodology** link to open the detailed pop-up with calculation methodology info.



Accuracy Validation Export

Learn more about [calculation methodology](#).

Search by data element or classifier

Data element/Classifier	Type	Validation date	Accuracy	Precision	Recall	
Cc Number	Data Element	01.11.2024, 05:42 PM	87%	99%	82%	⋮
SEXUAL ORIENTATION	Data Element	01.11.2024, 05:50 PM	100%	100%	100%	⋮
Cc Type	Data Element	01.11.2024, 06:11 PM	44%	63%	23%	⋮
Zip Code	Data Element	01.11.2024, 06:28 PM	90%	100%	55%	⋮
Domain	Data Element	-	-	-	-	⋮
Mothers Maiden	Data Element	-	-	-	-	⋮
Cc Expires	Data Element	-	-	-	-	⋮
Postcode	Data Element	-	-	-	-	⋮

Figure 3: Methodology link

PREREQUISITES FOR DATA ELEMENTS VALIDATION

To launch the data validation process for any of the data elements, you will need to upload your data set. For example, if you want to validate the credit card number data elements, prepare at least two TXT files. One of them should contain credit card numbers, and the other one should contain values that are not credit card numbers. These files should be then put inside a ZIP archive (maximum size of 5GB) with the following structure:

- 1 ZIP file with 2 folders;
- Folder 1 titled “Positive” (case insensitive) with credit card numbers. The minimum is 10 records, and the maximum is 100,000 records per file;
- Folder 2 titled “Negative” (case insensitive) with values that are not credit card numbers. The minimum is 10 records, and the maximum is 100,000 records per file;
- Every folder must contain at least 1 TXT file (maximum size of each TXT file 50MB). The system will recognize every row in the TXT file as a separate value, no separators needed. All Unicode characters and special values are allowed, and the maximum length per 1 row is 1024 characters.
- A maximum of 1000 files is allowed in total in the ZIP archive.

The ZIP archive being uploaded to IGDC should only include the necessary for processing files and folders at the root. The “Positive” and “Negative” folders must be located in the same directory (folder) before archiving. Any hidden/system files present in those folders and the archive itself can cause the upload to fail.

On MacOS:

Use one of the following commands to create a clean ZIP:

```
zip -r archive.zip . -x '**/*.*' -x '**/__MACOSX'
```

Or set this variable before zipping or tarring:

```
export COPYFILE_DISABLE=1
```

On Windows:

Before zipping, make sure the “Positive” and “Negative” folders are in the same directory. Select them (not the parent folder), then compress.

PREREQUISITES FOR CLASSIFIER VALIDATION

To launch the data validation process for any of the classifiers, you will need to upload your data set. The data set will be different depending on whether you want to validate an image classifier (supported for credit card, passport, and driver's license images) or a text-based classifier.

For example, if you want to validate the credit card image classifier, prepare at least two image files. One of them should be an image of a credit card, and the other one should be an image of something else. These files should be then put inside a ZIP archive (maximum size of 5GB) with the following structure:

- 1 ZIP file with 2 folders;
- Folder 1 titled “Positive” (case insensitive) with credit card images (PNG, JPEG (JPG), TIFF, BMP, GIF);
- Folder 2 titled “Negative” (case insensitive) with non-credit card images (PNG, JPEG (JPG), TIFF, BMP, GIF);
- Every folder must contain at least 1 image file (maximum size of each file 50MB);.
- A maximum of 1000 files is allowed in total in the ZIP archive.

As another example, if you want to validate a vacation request classifier , prepare at least two TXT files. One should contain values related to the vacation request fingerprint, and the other with values that are not

related to the fingerprint. These files should be then put inside a ZIP archive (maximum size of 5GB) with the following structure:

- 1 ZIP file with 2 folders;
- Folder 1 titled “Positive” (case insensitive) with related values. The minimum is 10 records, and the maximum is 100,000 records per file;
- Folder 2 titled “Negative” (case insensitive) with unrelated values. The minimum is 10 records, and the maximum is 100,000 records per file;
- Every folder must contain at least 1 TXT file (maximum size of each TXT file 50MB). The system will recognize every row in the TXT file as a separate value, no separators needed. All Unicode characters and special values are allowed, and the maximum length per 1 row is 1024 characters.
- A maximum of 1000 files is allowed in total in the ZIP archive.

The ZIP archive being uploaded to IGDC should only include the necessary for processing files and folders at the root. The “Positive” and “Negative” folders must be located in the same directory (folder) before archiving. Any hidden/system files present in those folders and the archive itself can cause the upload to fail.

On MacOS:

Use one of the following commands to create a clean ZIP:



```
zip -r archive.zip . -x '**/.*' -x '**/__MACOSX'
```

Or set this variable before zipping or tarring:

```
export COPYFILE_DISABLE=1
```

On Windows:

Before zipping, make sure the “Positive” and “Negative” folders are in the same directory. Select them (not the parent folder), then compress.

DATA VALIDATION

Perform the following steps to launch the data validation process.



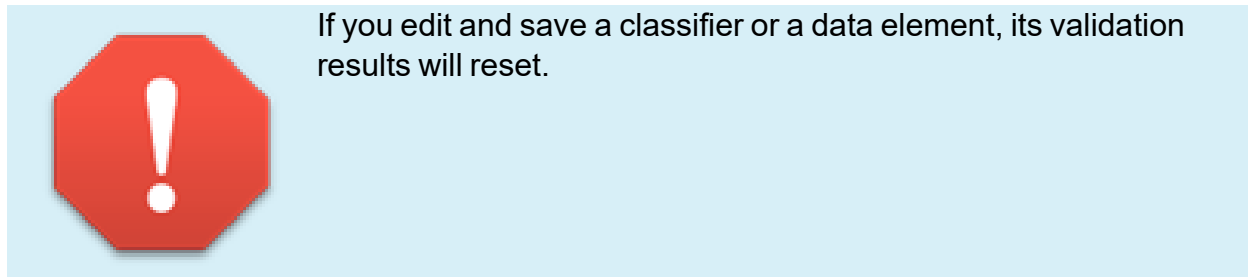
1. Click the three-dot button next to a specific data element/classifier to open its additional options menu.

Data element/Classifier	Type	Validation date	Accuracy	Precision	Recall	
Cc Number	Data Element	01.11.2024, 05:42 PM	87%	99%	82%	⋮
SEXUAL ORIENTATION	Data Element	01.11.2024, 05:50 PM	100%	100%	100%	⋮
Cc Type	Data Element	01.11.2024, 06:11 PM	44%	63%	23%	⋮
Zip Code	Data Element	01.11.2024, 06:28 PM	90%	100%	55%	⋮
Domain	Data Element	-	-	-	-	⋮
Mothers Maiden	Data Element	-	-	-	-	⋮

Provide new data set
 Change data element configuration

Figure 4: Additional options

2. Select the **Change data element/classifier configuration** option to go to the edit page for the corresponding data element/classifier (see the **Analytic Engine & Console Manager Admin Guide** for details on how to edit data elements/classifiers). Select the **Provide new data set** option to open the **Data sets upload** page.



Provide new data set

Upload data sets for accuracy validation

To validate data classification accuracy, please submit reference data samples for both True Positive and True Negative datasets. Please upload 1 zip file containing these as 2 separate folders, each folder must contain at least 1 file. Please note that accuracy results are directly dictated by Data element configuration.

- The minimum requirement for True Positive/True Negative samples is 10 records, with a maximum of 100,000 records per file (data element) and a maximum of 1000 files.
- Accepted file formats (content within zip file): .txt.
- File size limitation: 5 GB.

Please note, the accuracy validation is done in memory and the data sets will not be stored. [Download an example](#) how file should be formatted.

Data set*



Click or drag ZIP file here to upload

Cancel Validate

Figure 5: Upload page for data elements

Upload data sets for accuracy validation

To validate data classification accuracy, please submit reference data samples for both True Positive and True Negative datasets. Please upload 1 zip file containing these as 2 separate folders, each folder must contain at least 1 file. Please note that accuracy results are directly dictated by Document Classification Configuration.

- The minimum requirement for True Positive/True Negative samples is 2 files, with a maximum of 1000 files (fingerprints, images).
- Accepted file formats (content within zip file): .txt, image.
- File size limitation: 5 GB.

Please note, the accuracy validation is done in memory and the data sets will not be stored. [Download a text example](#) or [Download an image example](#) how file should be formatted.

Data set*



Click or drag ZIP file here to upload

Cancel Validate

Figure 6: Upload page for classifiers

3. Click the upload pane or drag and drop a ZIP file with true positive and true negative data sets as two separate files within the archive.

4. Click the **Validate** button to go back to the **Data accuracy** page and begin the file upload and data validation.

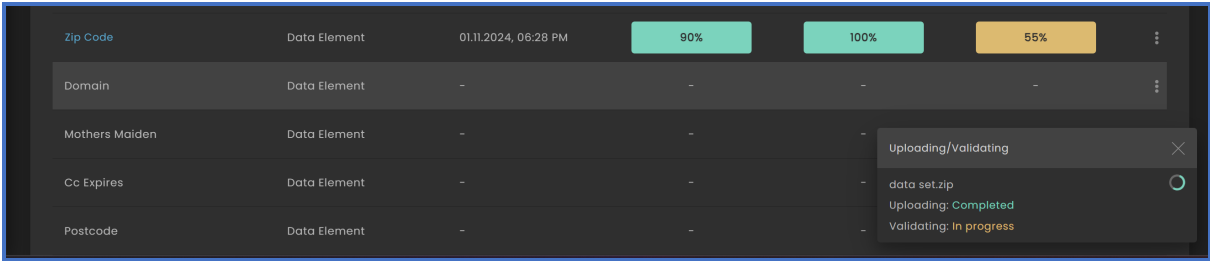


Figure 7: Upload process



Make sure to perform any action in the IGDC system at least every 15 minutes to avoid automatic log out as exiting the system will reset the upload and validation process.

When the upload process successfully ends and validation begins, it will no longer be possible to cancel it until validation is completed.

ACCURACY REPORT EXPORT

1. Once the validation process for the desired data element/classifier is completed, its title on the Data accuracy page becomes clickable. Clicking it will open the detailed statistics information for this specific data element or classifier.



Keep in mind that IGDC does not save data validation history, only the last result.

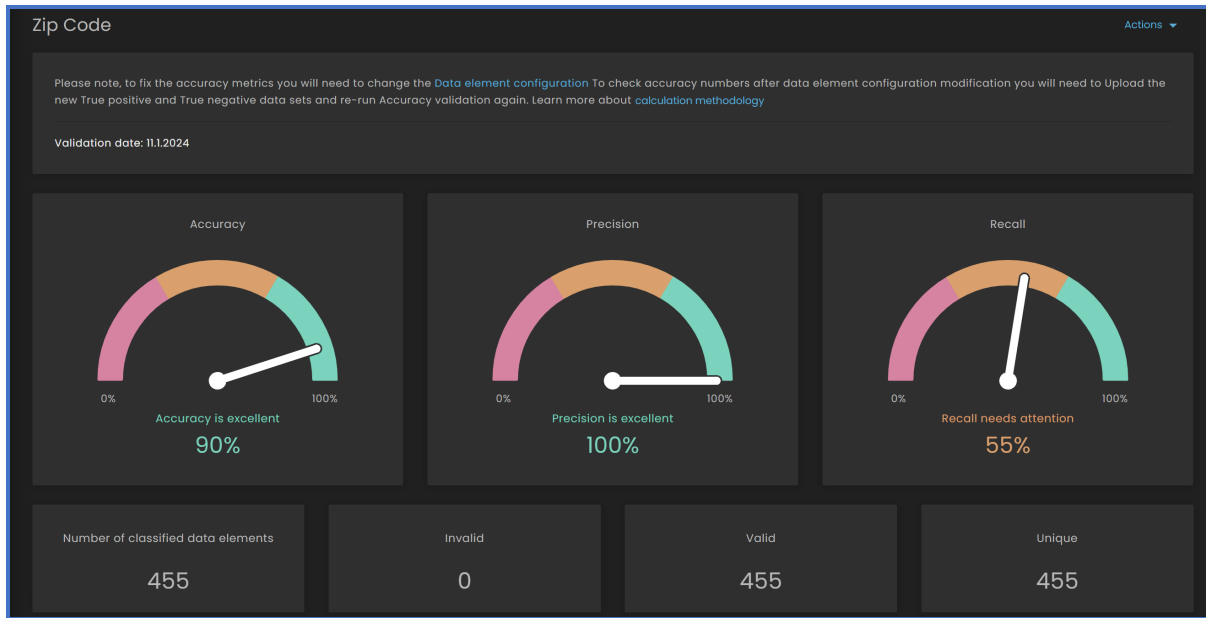


Figure 8: Validation results

Table 9: Properties on the detailed information page

PROPERTY	DESCRIPTION
Validation date	The date and time of the last successful validation.
Accuracy	Percentage of correctly classified instances out of the total instances. Accuracy rate: $(\text{True Positive} + \text{True Negative}) / (\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative})$
Precision	Percentage of positive predictions that are actually correct. Precision = $\text{True Positive} / (\text{True Positive} + \text{False Positive})$
Recall	Represents the model's ability to find all the positive instances. Recall rate: $= \text{True Positive} / (\text{True Positive} + \text{False Negative})$
Number of classified data elements	Shows the number of records used in accuracy validation
Invalid	Shows the number of invalid records used in accuracy validation (for example, an empty row or a broken character)
Valid	Shows the number of valid data set records used in accuracy validation (those that are not an empty row and do not have any broken characters)
Unique	Shows the number of unique data set records used in accuracy validation
True positive	The number of correctly classified records. Example: the number of correctly classified credit card numbers that actually are credit cards
True negative	The number of correctly classified records. Example: The number of non-credit card numbers correctly classified as non-credit card numbers
False positive	Amount of records mistakenly classified as an item belonging to a certain category when, in fact, it does not. Example: the number of non-credit card numbers incorrectly classified as credit card numbers
False negative	Amount of records mistakenly classified as an item not belonging to a certain category when, in fact, it does.

PROPERTY	DESCRIPTION
	Example: the number of credit card numbers incorrectly classified as non-credit card numbers

2. Go to **Actions > Export** to export the data element validation results as XLSX or CSV.

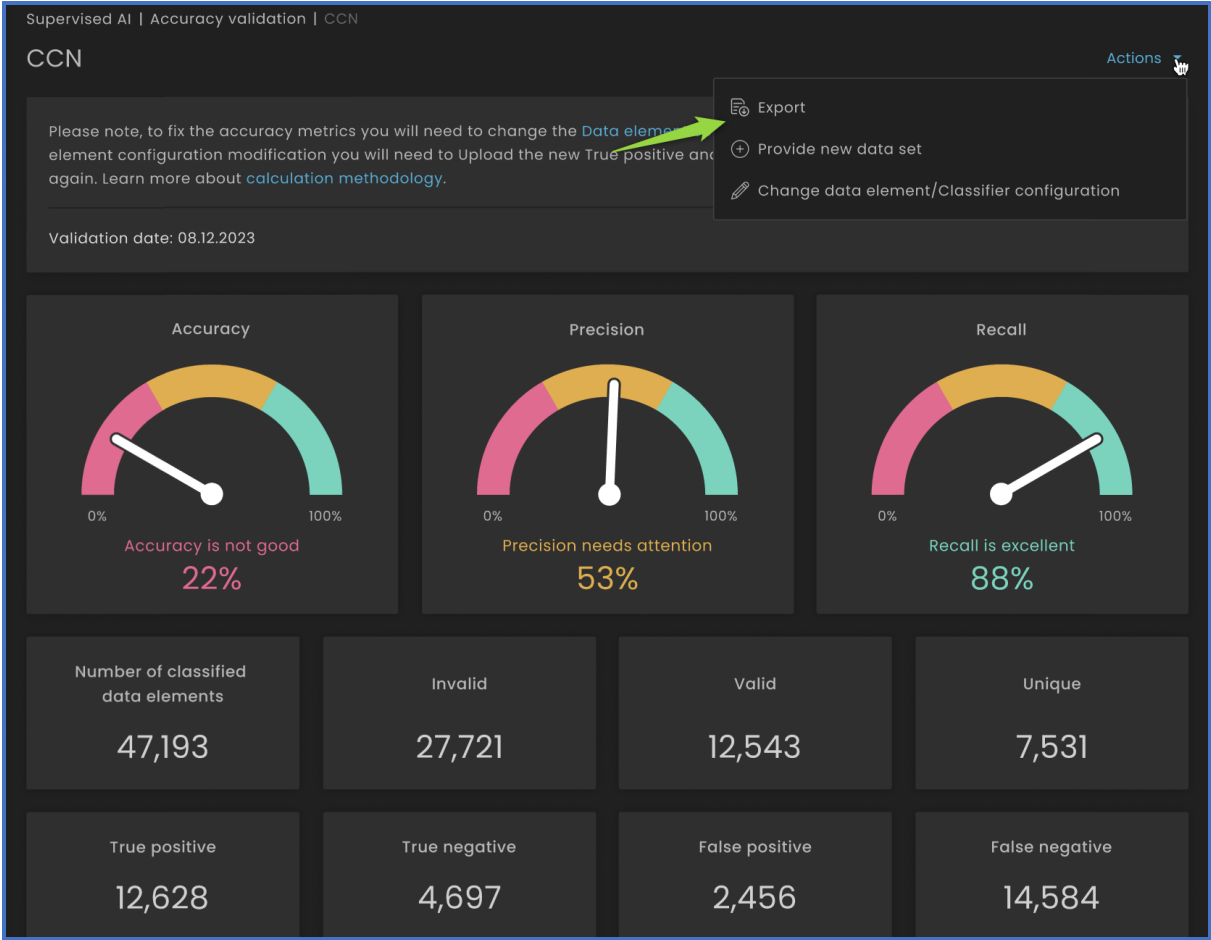


Figure 10: Export the results of data element validation

3. To export all data elements and classifier validation results go back to the **Data accuracy** page and click the **Export** button to export the data discovery and classification accuracy report (XLSX or CSV).

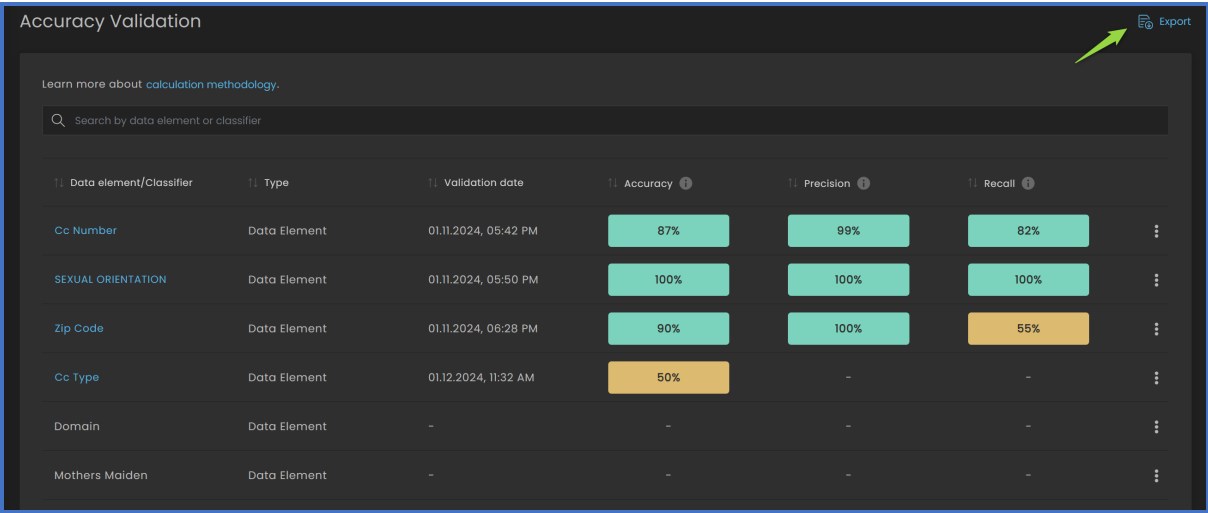


Figure 11: Export the full accuracy report

CHAPTER 3: CATALOG RECREATION

This chapter explains how to initiate the master catalog recalculation, select the required recalculation option, and review the recalculation history.

CATALOG RECALCULATION



Only users with SUPER_ADMIN permissions has access to the **Re-calculation** page. For other roles (admin, user, DSR operator), the **Re-calculation** option is not available in the sidebar menu.

1. To re-create the master catalog, go to **Sidebar > Settings > Data Quality > Re-calculation**. IGDC offers two options of master catalog recreation: **Re-create** or **Re-create & Drop**.
2. Click the **Re-create** or **Re-create & Drop** button to run the re-calculation. Select the appropriate option using the table below.

Table 1: Re-calculation Options

	RE-CREATE	RE-CREATE & DROP
Reason for re-calculation	* RDA has been modified or deleted. * Data element configuration has been changed: data element recognition settings (recognition rules, uniqueness) or constraints. * Master catalog preparation for production if it contains test and redundant data after the data quality test. * Upgrade to higher version involves data migration and requires the master-catalog recreation.	* Master catalog must contain only trusted data, excluding the automatically discovered data values.
Re-calculation process	Data sources analysis is active. Retrieved data processing is stopped. IGDC applies new Data element configurations and RDAs to the master catalog.	Data sources analysis is active. Retrieved data processing is stopped. IGDC applies new Data element configurations and RDAs to the master catalog, and removes all the automatically discovered values from the catalog.
Re-calculation result	Master catalog with <u>trusted</u> and <u>discovered</u> values according to the current Data element configuration and RDA settings.	Master catalog with <u>only trusted</u> values according to the current Data element configuration and RDA settings.

Settings / Data Quality / Re-Calculation

Data Quality Recalculation

Last re-creation:

Inventa supports recalculation of the Master Catalog at any time without affecting the overall discovery and classification effort. When the system recalculates the catalog, it will reload all the Root Data Assets (RDAs) from scratch (not just delta changes) and will re-create the Data Subjects Catalog again. **Please note** that before running the Master Catalogue recreation you have uploaded the latest RDA and have configured the data elements configuration.

Please note that the recalculation process will not involve analysis and will not affect your databases. After a successful recalculation, the system will return to the data discovery process.

Re-create Data Subjects Catalog

Re-create the Master Catalog according to the actual RDAs to visualize the relevant Trusted data. Please note that the historically discovered information will continue to appear in Data Subject Profile.

RE-CREATE

Re-create Data Subjects Catalog & Drop Discovered Data Element Types

Re-create the Master Catalog according to the actual RDAs and leave only Trusted data to appear in Data Subject Profile. All historically discovered data will be deleted and the Data Subject profiles will be re-inventoried according to the new RDAs and configurations.

RE-CREATE & DROP

Figure 2: Data quality recalculation page

3. Once you click **Re-create** or **Re-create & Drop**, the system will show a warning popup. Click **Yes** to initiate recalculation or click **Cancel** to exit without changes to the master catalog.



If you have no root data assets uploaded and run the recalculation, all your progress on discovered personal data will be lost during recalculation.

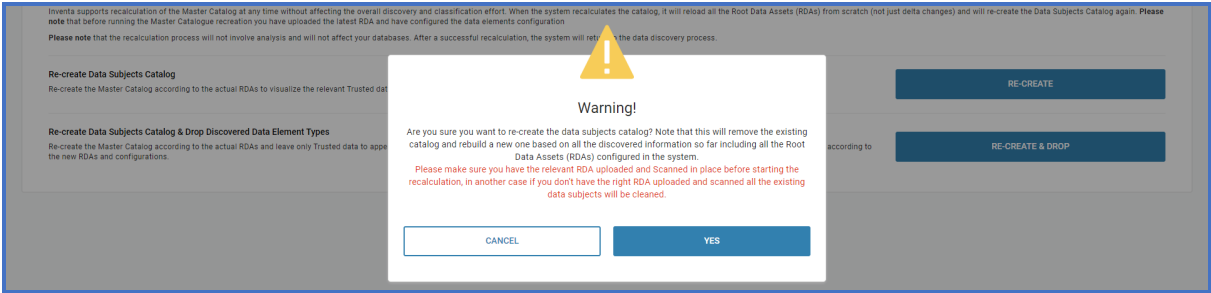


Figure 3: Warning popup on master catalog re-creation

4. Once you click **Yes**, IGDC will apply new data recognition settings according to the selected re-creation option. The page will show recalculation progress and the corresponding notification in the upper right corner:



PARAMETER	DESCRIPTION
Progress bar (1)	Master catalog re-creation progress in percent of analyzed records.
Total Record Received (2)	Number of records (values) in the master catalog.
Total Record Updated (2)	Number of records (values) affected during the master catalog re-creation.

To stop the re-creation process, click the **Stop** button (3). The process will stop immediately, and the



system will show the notification in the upper right corner: . The event will appear in the recalculation history.

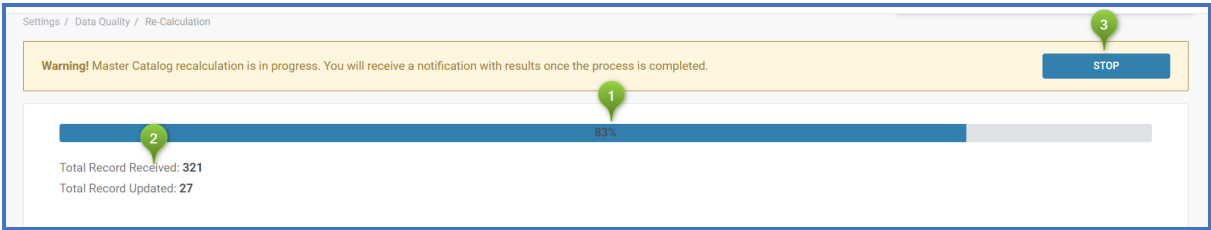


Figure 4: Master catalog re-creation progress

5. After the re-creation completion, IGDC will redirect you to the Re-Calculation page, the master catalog will be updated according to the selected option, and the event will appear in the recalculation history.

RECALCULATION HISTORY

To review the history of the master catalog re-creation, go to **Sidebar > Settings > Data Quality > Re-calculation history**.

Table 1: Re-calculation history parameters

PARAMETER	DESCRIPTION
Recalculation type	Type of the master catalog re-creation. Options: Recreate Master Catalog / Recreate Master Catalog & Drop.
Recalculation status	Current state of the master catalog re-creation. Options: Active / Finished.
Recalculation status	Date and time when the master catalog re-creation started.
Recalculation finished	Date and time of the master catalog re-creation completion. Available only for events with status = Finished.
Total records received	Number of records (values) in the master catalog.
Total records updated	Number of records (values) affected during the master catalog re-creation.
Total records removed	Number of records (values) removed from the master catalog due to RDA deletion.
Number of Data Subjects before recalculation	Number of data subject's profiles in the master catalog before re-creation.
Number of Data Subjects after recalculation	Number of data subject's profiles in the master catalog after re-creation completion.

Settings / Data Quality / Re-Calculation History								
Recalculation type	Recalculation status	Recalculation started	Recalculation finished	Total records received	Total records updated	Total records removed	Number of Data Subjects before recalculation	Number of Data Subjects after recalculation
Recreate Master Catalog	Finished	22 Mar, 2022 - 22:57:12	22 Mar, 2022 - 22:57:14	666	78	0	30	30
Recreate Master Catalog	Finished	22 Mar, 2022 - 22:56:29	22 Mar, 2022 - 22:56:31	620	69	0	30	30
<< < > >> 10								

Figure 2: Re-calculation history page

IBM, the IBM logo, and IBM Guardium Discover and Classify are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.