



Data Quality Audit and Tools Comparison Report Summary



Prepared for

Health Data Research UK

Prepared by

Inspirata, Inc.

Contents

ABOUT INSPIRATA	1
OVERVIEW	1
PHASE 1: AUDIT DATASET METADATA	3
1.1 DATASET CATEGORIZATION	3
1.2 DATASET METADATA QUALITY AUDIT OUTCOMES.....	3
PHASE 2: EVALUATE DATA QUALITY TOOLS	5
1.1 PHASES OF DATA PROCESSING.....	5
1.2 KEY CAPABILITIES REQUIRED TO ADDRESS DATA QUALITY	6
1.3 DATA QUALITY TOOL CAPABILITIES	8
1.4 OPEN DATA QUALITY TOOL COMPARISON APPROACH.....	10
1.5 TOOLS AND SUPPORTED DATA SOURCE CONNECTIVITY AND FORMATS	13
1.6 OPEN DATA QUALITY TOOL COMPARISON MATRIX/RESULTS	16
PHASE 3: COMPARE DATA QUALITY PROFILING TOOLS.....	17
1.1 SYNTHETIC REFERENCE DATASET CREATION	17
1.2 DATA PROFILING TOOLS EVALUATED	18
1.3 DATA QUALITY DIMENSIONS	19
1.4 DATA PROFILING TOOLS COMPARISON APPROACH	19
1.5 DATA PROFILING TOOLS COMPARISON SUMMARY RESULTS	23
1.6 THE DATA QUALITY PROCESS	24
1.7 SKILLS NEEDED FOR DATA PROFILING	24
OPEN DATA QUALITY PROFILING TOOLS OVERVIEW, FEATURES, STRENGTHS & LIMITATIONS	25
1.1 KNIME	26
1.2 ORANGE.....	26
1.3 PANDAS PROFILING (PYTHON).....	26
1.4 WEKA.....	26
1.5 AGGREGATE PROFILER	26
1.6 DATACLEANER.....	26
1.7 TALEND OPEN STUDIO FOR DATA QUALITY	26
1.8 MOBYDQ	26
INFO CARDS - OTHER OPEN SOURCE DATA QUALITY TOOLS	26
CONCLUSION	27
RECOMMENDATIONS.....	28

About Inspirata

Inspirata is a revolutionary healthcare IT company, using cutting edge technology to help organisations and patients realize more value from healthcare data. Our goal is to transform care by empowering health systems and providers with the real-time, system-wide data and intelligence needed to improve care quality and the patient experience, and ultimately, the economics of health and wellness.

Inspirata combines the power of an open technology platform, natural language processing and AI, and collaborative clinical applications to bring together disparate patient data and transform it into intelligence. We enable our customers and partners to take advantage of the platform to innovate—rapidly generating a new era of applications to improve healthcare and deliver superior outcomes to patients. Our goal is to make it easy for caregivers across the entire healthcare continuum to gain the insight they need to collaborate and provide the best patient care possible.

Inspirata's Clinical Data Platform provides healthcare institutions the ability to bring all their data together into one place. From this megastore, numerous applications can be built, such as: analytical and data mining engines, customized workflow management, and simplified sharing of critical information with physicians and other healthcare providers worldwide.

Our industry-leading oncology specific Natural Language Processing (NLP) technology provides value by automating manual data curation processes across the oncology service line including cancer case-finding, cancer reporting, clinical trial candidate identification and quality monitoring.

Inspirata's solutions enable our customers to achieve efficiencies, delivery higher quality care, re-duce costs, engage in ground-breaking research programs, and participate in partnerships with pharma/biotech companies.

Overview

HDR UK's mission is to unite the UK's health data to enable discoveries that improve people's lives. It is their ambition to provide a consistent national view on the qualities of any dataset through their Innovation Gateway. This would allow users to understand whether a dataset may meet their needs, ahead of requesting access.

It is important to understand that data quality is a component of data utility and the same data resource may have an acceptable level of quality for some contexts, but this quality may be unacceptable for other contexts. Data quality metrics used in this project are derived impartially and disconnected from the specific contextual characteristics. These metrics allow for a form of benchmarking to be performed but may have poor relationship to overall utility. Inspirata assessed data quality along commonly used dimensions: - completeness, consistency, uniqueness, validity, accuracy, and timeliness.

The use of utility-driven assessment in real-world data management scenarios have broader implications and require a high degree of domain-specific subject matter expertise. It is therefore important to recognize that improving data quality across a wide range of health data types

requires significant investment and engagement of subject matter experts. It is also important to weigh the extent of data curation needed that would provide best value for money. Sharing best practices and learnings is important to ensure long term success in data quality management across the digital innovation hub sites and will ultimately generate learning for wider applicability and scalability. Semi-automated analysis and data quality profiling software tools only aid in the process.

The initial intent of the project was to establish a mechanism (framework) for the assessment of data quality. Although not in the original scope, the project, evolved into the generation of synthetic datasets and subsequent assessment of the quality of these datasets. In addition, the project established a level of understanding of the quality of datasets currently published on the HDR UK Innovation Gateway based on the metadata as well as an assessment of open data quality tools available in order to provide a recommendation to affiliate and alliance members.

The project consisted of three phases:

- **Phase 1** was to conduct an audit of the dataset metadata published on the HDR UK Innovation Gateway Metadata Catalogue.
- **Phase 2** involved the creation of a general comparison matrix of open data quality tools that can be used to evaluate and achieve data quality overall.
- **Phase 3** involved a detailed comparison of open data quality profiling tools.

Inspirata created three frameworks whereby future audits and updated comparisons may be made relating to:

1. Comparison of data quality tools
2. In-depth comparison of profiling tools
3. Comparison of meta-data

This document describes the approach, outcomes and recommendations, with additional supporting information about each of the data quality tools in appendices as follows:

- WhiteRabbit
- AggregateProfiler
- DataCleaner
- Knime
- MobyDQ
- Orange
- Pandas
- Talend Open Studio
- Weka

Phase 1: Audit dataset metadata

1.1 Dataset Categorization

Dr Varma, Director of Engineering, Health Data Research UK automated the process of extracting the datasets and can be found at <https://github.com/HDRUK/datasets>. The direct link to daily extract is <https://raw.githubusercontent.com/HDRUK/datasets/master/datasets.csv>

Inspirata would like to recognize the significant contributions of **Prof. Neil Sebire**. After meetings with Prof. Sebire to understand the data types and formats along with the typical problems and issues his team has faced, Inspirata was able to construct a framework whereby each dataset published on the HDR UK Innovation Gateway Metadata Catalogue (<https://metadata-catalogue.org/hdruk/#/catalogue/dataModel/all>) was categorized into a table.

DATASET CATEGORIES		DATA						
		1	2	3	4	5	6	7
MODEL & METADATA	A	1 Single Entity or Table / Structured File (e.g. JSON / XML / CSV / Single SQL table)	2 Relational Multi-Entity (SQL / Oracle)	3 Non-Relational Multi-Entity (NoSQL/Column) (e.g. FHIR/HL7/CCDA) / Multiple Files (e.g. CSVs)	4 Unstructured (e.g. images PACS/LIS Dicom)	5 Semi-Structured (e.g. clinical notes, diagnostic reports)	6 Qualitative (Questionnaires / CQL / QDM)	7 In progress / Not available / Unknown
	B	Standard Data Model & Metadata Provided/Unambiguous	A1	A2	A3	A4	A5	A6
	C	Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	B1	B2	B3	B4	B5	B6
	D	Proprietary/Non-Standard Data Model & Metadata Provided/Unambiguous	C1	C2	C3	C4	C5	C6
	E	Proprietary/Non-Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	D1	D2	D3	D4	D5	D6
	F	No Data Model & Metadata Provided/Unambiguous	E1	E2	E3	E4	E5	E6
	G	No Data Model & Metadata Not Provided/Ambiguous/Inadequate	F1	F2	F3	F4	F5	F6

Green = ideal, yellow = needs work, orange = limited, red = inadequate

1.2 Dataset metadata quality audit outcomes

Inspirata conducted two data quality audits on the metadata in April 2020 and in June 2020. We are happy to report an improvement in data quality overall and an increase in the total number of data sets available on the metadata catalog.

Number of data sets	APRIL 2020	JUNE 2020	
Total	415	446	↑
Green category	24	110	↑
Yellow category	27	1	↓
Orange category	17	18	
Red category	347	316	↓

The following tables show results from the dataset metadata quality audit.

APRIL 2020

TOTAL NUMBER OF HDR UK DATASETS: 415

HDR UK DATASET METADATA CATALOGUE (COUNT)		DATA						
		1	2	3	4	5	6	7
MODEL & METADATA	A	Standard Data Model & Metadata Provided/Unambiguous	1	0	0	0	0	
	B	Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	25	0	0	1	0	
	C	Proprietary/Non-Standard Data Model & Metadata Provided/Unambiguous	10	1	12	0	0	
	D	Proprietary/Non-Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	5	9	3	1	0	
	E	No Data Model & Metadata Provided/Unambiguous	1			0	0	
	F	No Data Model & Metadata Not Provided/Ambiguous/Inadequate	7			0	0	339

HDR UK DATASET METADATA CATALOGUE (%)		DATA						
		1	2	3	4	5	6	7
MODEL & METADATA	A	Standard Data Model & Metadata Provided/Unambiguous	0.24%	0.00%	0.00%	0.00%	0.00%	
	B	Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	6.02%	0.00%	0.00%	0.24%	0.00%	
	C	Proprietary/Non-Standard Data Model & Metadata Provided/Unambiguous	2.41%	0.24%	2.89%	0.00%	0.00%	
	D	Proprietary/Non-Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	1.20%	2.17%	0.72%	0.24%	0.00%	
	E	No Data Model & Metadata Provided/Unambiguous	0.24%			0.00%	0.00%	
	F	No Data Model & Metadata Not Provided/Ambiguous/Inadequate	1.69%			0.00%	0.00%	81.69%

JUNE 2020

TOTAL NUMBER OF HDR UK DATASETS: 446

HDR UK DATASET METADATA CATALOGUE (COUNT)		DATA						
		1	2	3	4	5	6	7
MODEL & METADATA	A	Standard Data Model & Metadata Provided/Unambiguous	2	0	0	1	2	
	B	Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	0	0	0	1	0	
	C	Proprietary/Non-Standard Data Model & Metadata Provided/Unambiguous	31	29	66	0	0	
	D	Proprietary/Non-Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	9	2	16	1	0	
	E	No Data Model & Metadata Provided/Unambiguous	0			0	0	
	F	No Data Model & Metadata Not Provided/Ambiguous/Inadequate	6			0	0	280

HDR UK DATASET METADATA CATALOGUE (%)		DATA						
		1 Single Entity / Structured File (e.g. JSON/XML/CSV)	2 Relational Multi- Entity (SQL)	3 Non-Relational Multi-Entity (NoSQL/Column) (e.g. FHIR/HL7) / Multiple Files (e.g. CSV)	4 Unstructured (e.g. images PACS/LIS Dicom)	5 Semi-Structured (e.g. clinical notes, diagnostic reports)	6 Qualitative (Questionnaires / CQL / QDM)	7 In progress / Not available / Unknown
MODEL & METADATA	A Standard Data Model & Metadata Provided/Unambiguous	0.45%	0.00%	0.00%	0.22%		0.45%	
	B Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	0.00%	0.00%	0.00%	0.22%		0.00%	
	C Proprietary/Non-Standard Data Model & Metadata Provided/Unambiguous	6.95%	6.50%	14.80%		0.00%	0.00%	
	D Proprietary/Non-Standard Data Model & Metadata Not Provided/Ambiguous/Inadequate	2.02%	0.45%	3.59%		0.22%	0.00%	
	E No Data Model & Metadata Provided/Unambiguous	0.00%			0.00%	0.00%	0.00%	
	F No Data Model & Metadata Not Provided/Ambiguous/Inadequate	1.35%			0.00%	0.00%	0.00%	62.78%

THE FOLLOWING TABLE SHOWS THE SOURCES AND DISTRIBUTION OF QUALITY OF DATASETS WE EVALUATED:

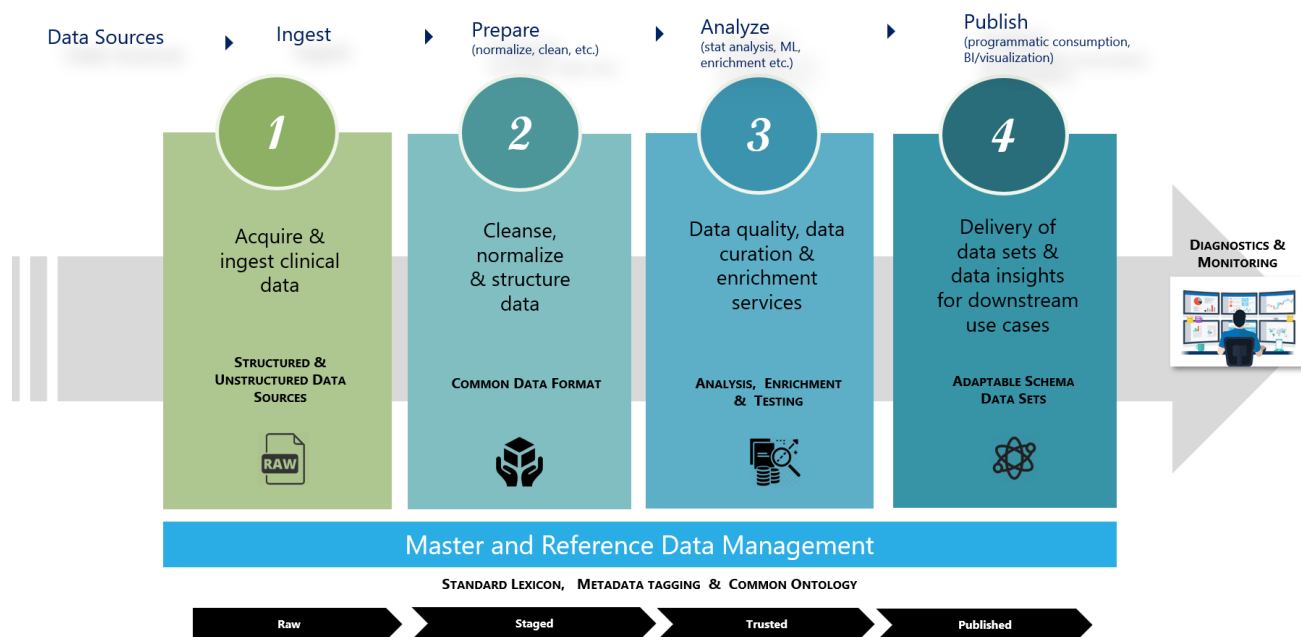
HDRUK Innovation Gateway Folder	A1	A4	A6	B4	C1	C2	C3	D1	D2	D3	D5	F1	F7	Grand Total
Barts				1							1		9	11
BREATHE									2	1		5	9	17
COG-UK	1													1
CPRD							35						1	36
Cystic Fibrosis										10				10
DATA-CAN							3							3
DISCOVER NOW					8		2							10
Dundee													23	23
Genomics England							6							6
Gut Reaction					3								1	4
HQIP													80	80
Insight							1						2	3
Leicester													3	3
NHS Digitrials					1		7					1		9
NHSX			1											1
NIHR BioResource	1		2		1								7	11
PHE							1							1
Pioneer													3	3
RCGP										1				1
SAIL						29	1							30
Scotland							7	4					10	21
Tissue Directory													105	105
HDR UK/NHS Digital					18		2	5		1			9	35
HDR UK/NIHR Health Informatics Collaborative							1			3				4
HDR UK/Oxford University Hospitals NHS Foundation Trust													17	17
HDR UK/Alspac													1	1
Grand Total	2	1	2	1	31	29	66	9	2	16	1	6	280	446

The supporting Excel document (*MetaData Catalogue Assessment June 2020.xlsm*) can be used to compare and rate additional data profiling tools as needed.

Phase 2: Evaluate data quality tools

1.1 Phases of data processing

As depicted in the diagram, there are many different aspects to achieving data quality as data transitions from a raw to a staged to a trusted and ultimately a published form ready for consumption. Different tools are used for different aspects.



1.2 Key capabilities required to address data quality

Using Gartner reporting as a guideline only (<https://www.gartner.com/en/documents/3913549>), Inspirata suggest the following key capabilities that organizations need in their tool portfolio, if they are to address the increasing importance and urgency of data quality:

Data Ingestion and Integration

- ✓ **Connectivity:** The capability to access, and apply data quality rules to, a wide range of data sources. These include internal and external data, on-premises and cloud data, and structured and unstructured data sources.
- ✓ **Parsing:** Built-in capabilities for decomposing data into its component parts.
- ✓ **Issue resolution and workflow:** The process flow and user interface that enables business users to identify, quarantine, assign, escalate and resolve data quality issues.
- ✓ **Architecture and integration:** Commonality, consistency and interoperability among the various components of the data quality toolset and third-party tools.

Data Preparation and Cleaning

- ✓ **Master Reference Data Management:** Master data management (MDM) is the process of creating one single master reference source for all critical business data, leading to fewer errors and less redundancy in business processes.
- ✓ **Standardization and cleansing:** Built-in capabilities for applying standards, business rules or knowledge bases to modify data for specific formats, values and layouts.
- ✓ **Matching, linking and merging:** Built-in capabilities for matching, linking and merging related data entries within or across datasets, using a variety of techniques, such as rules, algorithms, metadata and machine learning.
- ✓ **Address validation/geocoding:** Support for location-related data standardization and cleansing.
- ✓ **Data curation and enrichment:** The capability to integrate externally sourced data to improve completeness and add value.

Data Profiling, Exploration / Pattern Detection

- ✓ **Data profiling, measurement and visualization:** Data analysis capabilities that give business insight into the quality of data, and that help them identify and understand data quality issues.

Data Monitoring

- ✓ **Monitoring:** Capabilities to assist with the ongoing understanding and assurance of data quality by monitoring of, and alerting to, possible data quality issues.

Data Use

- ✓ **Metadata management:** The capability to capture, reconcile and interoperate metadata relating to the data quality process.
- ✓ **Usability:** Suitability of the tools to engage and support the various roles (especially business roles) required by a data quality initiative.
- ✓ **DevOps environment:** Capabilities that facilitate configuration of data quality operations.
- ✓ **Deployment environment:** Styles of deployment and hardware and operating system options for deploying data quality operations.

1.3 Data quality tool capabilities

The following are typical data quality tool capabilities and features:

TITLE	DESCRIPTION
Data Profiling	<i>The process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.</i>
Relationship discovery	Discovering how parts of the data are interrelated. For example, key relationships between database tables, references between cells or tables in a spreadsheet. Understanding relationships is crucial to reusing data; related data sources should be united into one or imported in a way that preserves important relationships
Content discovery	Looking into individual data records to discover errors. Content discovery identifies which specific rows in a table contain problems, and which systemic issues occur in the data (for example, phone numbers with no area code)
Structure discovery	Validating that data is consistent and formatted correctly and performing mathematical checks on the data (e.g. sum, minimum or maximum). Structure discovery helps understand how well data is structured—for example, what percentage of phone numbers do not have the correct number of digits
Data Integration	<i>The process of combining data from several disparate sources and providing a unified view of the data.</i>
Data Consolidation	Data consolidation physically brings data together from several separate systems, creating a version of the consolidated data in one data store. Often the goal of data consolidation is to reduce the number of data storage locations.
Data Propagation	Data propagation is the use of applications to copy data from one location to another
Data Virtualization	Provide a near real-time, unified view of data from disparate sources with different data models. Data can be viewed in one location but is not stored in that single location.
Data Federation	Federation is technically a form of data virtualization. It uses a virtual database and creates a common data model for heterogeneous data from different systems. Data is brought together and viewable from a single point of access.

Data Cleaning	<i>After determining that a data value does not conform to end-user expectations data can be transformed into a form that meets the level of business user acceptability. Data cleansing builds on the parsing, standardization, and enhancement tools; identity resolution; and record linkage. By parsing the values and triggering based on known error patterns, data cleansing will apply rules to figure out the right data values, correct names or addresses, eliminate extra bits of information, reduce meaningless data, and even merge duplicates.</i>
Parsing and Standardization	Data values are expected to conform to expected formats and structures, but slight variations in data values may confuse individuals or automated applications. Parsing is used to determine whether a value conforms to recognizable patterns. When patterns are recognized, other rules and actions can be triggered to transform the input data into a form that can be more effectively used, either to standardize the representation (presuming a valid representation) or to correct the values
Identity Resolution, Linkage, Merging and Consolidation of Duplicate Records	Identity resolution is used to recognize when only slight variations suggest that different records are connected and where values may be cleansed, or where enough differences between the data suggest that the two records represent distinct entities. Identity resolution provides the foundation for duplicate record analysis and elimination. Identifying similar records within the same data set means that the records may be subjected to cleansing, elimination, or both. Identifying similar records in different sets may indicate a link across the data sets, which helps facilitate merging of similar records for the purposes of data cleansing as well as supporting a master data management or customer data integration initiative. Automating the merging process will select the best data values and allow for the creation of a single “best copy”.
Master Reference Data Management	Master data management (MDM) is the process of creating one single master reference source for all critical business data, leading to fewer errors and less redundancy in business processes.
Data Use	<i>Enables users to gain useful insights from the available data</i>
Metadata Management	Metadata gives context, helping to organize and provide relevance to the data itself. Attributes like file location, file size, data type, and author are vital signposts that allow for faster querying and replication of insights. Enabling a business user to search and identify the information on the key attributes in web-based user interface. Business users can understand where the data for an attribute is coming from and how the data in the attribute was calculated. They can visualize which enterprise systems in the organization the attribute is being used in (Lineage) and understand the impact of changing something (Impact Analysis) related to the attribute such as the length of the attribute to other systems.
Privacy & Security	DevOps capabilities for configuration and deployment and capabilities to anonymize/pseudonymous data and support the various roles (especially

	business roles) required by a data quality initiative without compromising HIPAA/GDPR requirements
Data Mining	Process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database system
Data Monitoring	<i>Capabilities to assist with the ongoing understanding and assurance of data quality by monitoring of, and alerting to, possible data quality issues</i>
Database analysis	Data mining algorithms to extract information from various heterogeneous databases and anticipate evolving patterns
Data Availability	Identify incomplete, stale or missing data
Data Visualization & Exploration	Understand patterns, trends, and distributions with scatter plots, histograms, line charts, parallel coordinates, box plots, and find and fix common data quality problems including missing values and outliers

1.4 Open data quality tool comparison approach

Open Data Quality Tools were compared and categorized based on available product documentation and data sheets as well as experience with these and similar commercial tools. Each tool was ranked based on key capabilities required to address data quality using the following feature tree and scoring:

Summary of Features and Possible Scores per Feature

Feature	Possible Score
Data Profiling, Exploration/Pattern Detection	45
Data Ingestion and Integration	47
Data Preparation and Cleaning	90
Data Use	80
Data Monitoring	15

Detailed Scoring Table per Feature

Feature	Score	Category
Data Profiling, Exploration/Pattern Detection	45	
Relationship discovery		
Cross Table Redundancy Analysis	5	Data profiling, measurement, visualization
Performing data quality assessment, risk of performing joins on the data	5	Data profiling, measurement, visualization
Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded	5	Data profiling, measurement, visualization

value dependencies, and performing inter-table analysis.		
Content discovery		
Data Pattern Discovery	5	Data profiling, measurement, visualization
Domain Analysis	5	Data profiling, measurement, visualization
Discovering metadata and assessing its accuracy	5	Data profiling, measurement, visualization
Structure discovery		
Column Value Frequency Analysis & Statistics, collecting descriptive statistics like min, max, count and sum.	5	Data profiling, measurement, visualization
Table Structure Analysis, Collecting data types, length and recurring patterns.	5	Data profiling, measurement, visualization
Drill-through Analysis	5	Data profiling, measurement, visualization
Data Ingestion and Integration	47	
Data Consolidation		
Connectivity to < 3 data sources	3	Connectivity
Connectivity to < 6 data sources	4	Connectivity
Connectivity to > 5 data sources	5	Connectivity
(ETL) Data Extraction, Transformation and Loading / ETL and ELT support	5	Parsing
Data Modelling	5	Architecture & integration
Data Propagation		
Data flow orchestration, Enterprise application integration (EAI), exchange of messages and transactions	5	Issue resolution & workflow
Enterprise data replication (EDR), transfer large amounts of data between databases	5	Architecture & integration
Versioning and file management	5	Architecture & integration
Data Virtualization		
Data Access	5	Connectivity
Data Federation		
Enterprise information integration (EII)	5	Issue resolution & workflow
Data Preparation and Cleaning	90	
Parsing and Standardization		
Tagging data with keywords, descriptions or categories	5	Standardization & cleansing
Data Scrubbing/Cleansing/Handling blank values/Reformatting values/Threshold checking	5	Standardization & cleansing
Data Enhancement/Enrichment/Curation	5	Data curation & enrichment
NLP	5	Standardization & cleansing

Address validation/geocoding	5	Address validation / geocoding
Master Data Management	5	Standardization & cleansing
Data masking	5	Data curation & enrichment
Identity Resolution, Linkage, Merging & Consolidation		
Data Deduping	5	Matching, linking & merging
Machine Learning / Training a statistical model	5	Data curation & enrichment
Data aggregation	5	Matching, linking & merging
Data Binning	5	Matching, linking & merging
Grouping similar data / Clustering	5	Matching, linking & merging
Outlier detection and removal	5	Matching, linking & merging
Master Reference Data Management		
"Hub" infrastructure to source and distribute master/reference data	5	Master Reference Data Management
Master data versioning based on data history and timelines	5	Master Reference Data Management
Workflow integrations to steward and publish the master/reference data	5	Master Reference Data Management
Graph data stores to define relationships for creating a flexible knowledge graph	5	Master Reference Data Management
Accessible API for real-time access to shared reference data	5	Master Reference Data Management
Data Use	80	
Metadata Management		
Concept Identification and Naming	5	Metadata management
Data Categorization	5	Metadata management
Lineage	5	Metadata management
Relationship with other metadata	5	Metadata management
Comments and Remarks	5	Metadata management
Data Stats (profiles)	5	Metadata management
Knowledge Graph	5	Metadata management
Privacy & Security		
Data Anonymization	5	Usability
Role based access control	5	Usability
Secure environment setup and deployment	5	DevOps environment
Container based deployment	5	Deployment environment
Data Mining		
Interactive Data Visualization	5	Usability
Visual Programming and analysis	5	Usability
Visual Illustrations and training documentation	5	Usability
Sample Data / Generate Fake Data	5	Usability
Add-ons and Extension Functionality	5	Architecture & integration

Data Monitoring	15	
Monitoring & Alerting		
Time series data identified and collection by metric name and key/value pairs	5	Monitoring
Flexible query language to leverage this dimensionality	5	Monitoring
Graphing and dashboarding support	5	Monitoring

1.5 Tools and Supported Data Source Connectivity and Formats

The following is a list of the tools that were compared:

Tool	Connectivity	Data Sources / File Formats
Knime (Data analytics, profiling, reporting and integration platform)	Connectivity to > 5 data sources	Simple text formats (CSV, PDF, XLS, JSON, XML, etc.)
		Unstructured data types (images, documents, networks, molecules, etc.)
		Time series data
		Connect to a host of databases and data warehouses to integrate data from Oracle, Microsoft SQL, Apache Hive, and more
		Load Avro, Parquet, or ORC files from HDFS, S3, or Azure
		Access and retrieve data from sources such as Twitter, AWS S3, Google Sheets, and Azure and extended via pandas
Pandas Profiling (using Pandas I/O) (Python module for exploratory data analysis (EDA))	Connectivity to > 5 data sources	Text: - CSV, fixed-width text files, JSON, HTML, Clipboard, Excel
		Binary: OpenDocument, HDF5 Format, Feather Format, Parquet Format, ORC Format, Msgpak, Stata, SAS, SPSS, Python Pickle Format
		SQL, Google BigQuery
Orange (Data visualization, machine learning, data profiling and mining toolkit)	Connectivity to > 5 data sources	Excel (.xlsx), simple tab-delimited (.txt), comma-separated files (.csv) or Google Sheets document
		distance matrix: Distance File
		predictive model: Load Model
		network: Network File from Network add-on
		images: Import Images from Image Analytics add-on
		several spectroscopy files: Multifile from Spectroscopy add-on
RapidMiner (LIMITED FREE VERSION) (Integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics)	Connectivity to > 5 data sources	PostgreSQL, SQL, online repository, and extended via pandas
		Files: CSV, Stata, Hyper (Tableau), XLS, XML, QlikView, and more
		SQL: AccessDB, HSQLDB, Microsoft SQL Server (JTDS / Microsoft), MySQL, Oracle, PostgreSQL, Sybase
		NoSQL: Cassandra, MongoDB, Solr, Splunk (read only)
WEKA (Machine learning software to solve data mining problems)	Connectivity to < 3 data sources	Cloud services: Amazon S3, Azure blob and data lake, Dropbox, Google, Salesforce, Twitter, Zapier, Salesforce
		Arff, JSON, CSV, xrrf, dat, data, names, and more
		Database using ODBC

Anonimatron (Pseudonymizes datasets)	Connectivity to > 5 data sources	Oracle, PostgreSQL, MySQL, DB2, MsSQL, Cloudscape, Pointbase, Firebird, IDS, Informix, Enhydra, Interbase, Hypersonic, jTurbo, SQLServer and Sybase
ARX Data Anonymization (Scalable Data Anonymization Tool - supports multiple privacy models)	Connectivity to > 5 data sources	CSV files, MS Excel spreadsheets Relational database systems, such as MS SQL, DB2, MySQL or PostgreSQL
WhiteRabbit (Tool to help prepare for ETLs of healthcare datasets)	Connectivity to > 5 data sources	comma-separated text files MySQL, SQL Server, Oracle, PostgreSQL, Microsoft APS, Microsoft Access, Amazon RedShift, Google BigQuery
Aggregate Profiler (AP) (Data profiling and analysis tool)	Connectivity to > 5 data sources	XML, XLS or CSV format, PDF export Teiid, Mysql, Oracle, Postgres, Access, Db2, SQL Server certified Big data support - HIVE
Talend Open Studio for Data Integration (LIMITED FREE VERSION) (Data integration and ETL)	Connectivity to > 5 data sources	More than 900 pre-built connectors and components for Oracle, Teradata, Microsoft SQL server, Marketo, Salesforce, NetSuite, SAP, Microsoft Dynamics, Sugar CRM, Dropbox, Box, SMTP, FTP/SFTP, LDAP, and more
Talend Open Studio for Big Data (LIMITED FREE VERSION) (ETL for large and diverse data sets)	Connectivity to > 5 data sources	Cloud: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and more RDBMS: Oracle, Teradata, Microsoft SQL server, and more SaaS: Marketo, Salesforce, NetSuite, and more Packaged Apps: SAP, Microsoft Dynamics, Sugar CRM, and more Technologies: Dropbox, Box, SMTP, FTP/SFTP, LDAP, and more
Talend Open Studio for Data Quality (LIMITED FREE VERSION) (Assesses accuracy and integrity of data - Data Profiling Tool)	Connectivity to > 5 data sources	Local or remote file that can be imported into the Talend Data Preparation tool (or from a database connection or other data sources, although not in the context of the Free Desktop version). Excel or CSV file 90+ data sources and scale with Stitch Data Loader - https://www.talend.com/products/pricing-model/
Talend Open Studio for ESB (LIMITED FREE VERSION)	Connectivity to > 5 data sources	Cloud: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and more RDBMS: Oracle, Teradata, Microsoft SQL server, and more SaaS: Marketo, Salesforce, NetSuite, and more Packaged Apps: SAP, Microsoft Dynamics, Sugar CRM, and more Technologies: Dropbox, Box, SMTP, FTP/SFTP, LDAP, and more
Talend Open Studio for MDM (LIMITED FREE VERSION) (key capabilities for data governance and master data management)	Connectivity to > 5 data sources	AWS, Microsoft Azure, Google Cloud Platform, and more. Plus, SaaS, packaged apps, and web services
OpenRefine (Tool for cleaning and transforming data)	Connectivity to < 3 data sources	TSV, CSV, *SV, .xls, .xlsx, JSON, XML, RDF as XML and google documents
DataCleaner		CSV files, Excel spreadsheets

(COMMUNITY EDITION - Limited) (Data profiling, data cleaning, and data integration tool) - offers integration with Pentaho	Connectivity to > 5 data sources	JDBC, MySQL, PostgreSQL, SQL Server Salesforce, SugarCRM
DataPreparator (Preprocessing - data cleaning, transformation, and exploration)	Connectivity to < 3 data sources	JDBC, XLS ARFF, DATA, CSV or plain text file format
Data Match (30-DAY FREE TRIAL) (visual data cleansing application - a component of Data Ladder)	Connectivity to > 5 data sources	Access, Apache HBase, Dynamics CRM, Email, Excel, Facebook, JSON, MongoDB, MySQL, Salesforce, SugarCRM, Twitter, XML
DataMartist (30 DAY FREE TRIAL, STANDARD - \$349, PROFESSIONAL - \$995) (Visual, data profiling and data transformation tool)	Connectivity to > 5 data sources	SQL Server, Oracle, MySQL, ODBC, MS Access, Excel Spreadsheets, Delimited text files including CSV data
Pentaho Kettle (COMMUNITY EDITION - Limited) (ETL Tool) Integrates with WEKA (Data Profiling)	Connectivity to > 5 data sources	Oracle, PostgreSQL, Redshift, SAP, SQLite, SparkSQL, Sybase, Teradata, UniVerse, Verica, Cloudera Impala, Hypersonic, H2 and more
SQL Power Architect (COMMUNITY EDITION - Limited) (Data Modeling & Profiling Tool)	Connectivity to > 5 data sources	JDBC, PostgreSQL, SQL, MySQL, HSQLDB, Oracle, DB2, HSQLDB, SQLstream, H2, Derby
SQL Power DqGuru (COMMUNITY EDITION - Limited) (Data Cleansing & MDM Tool)	Connectivity to > 5 data sources	JDBC, Oracle, Postgress, MySQL, Sybase and more
DQ Analyzer (COMMUNITY EDITION - Limited) (Data profiling tool)	Connectivity to > 5 data sources	Oracle, MS SQL, DB2, Sybase, Teradata, MySQL, Apache Derby, PostgreSQL CSV, TXT, and XLS(X)
Pimcore (Data Management, Integration, PIM, MDM, DAM)		
CytoScape (software platform for visualizing molecular interaction networks and biological pathways)		Simple interaction file (SIF or .sif format), Graph Markup Language (GML or .gml format), XGMML (extensible graph markup and modelling language), SBML, BioPAX, PSI-MI Level 1 and 2.5, Delimited text, Excel Workbook (.xls)
Anaconda (data science platform)	Connectivity to > 5 data sources	Multiple Python Connectors
Pyxplorer (a simple tool that allows interactive profiling of datasets)	Connectivity to < 5 data sources	Hive, Impala, MySQL

MobyDQ (Testing tool - aims to automate Data Quality checks during data processing)	Connectivity to > 5 data sources	Cloudera Hive, MariaDB, Microsoft SQL Server, MySQL, Oracle, PostgreSQL, SQLite, Teradata, Snowflake, Hortonworks Hive
---	----------------------------------	--



1.6 Open data quality tool comparison matrix/results

Tools were categorized, features logged and ultimately scored using the scoring weight shown in the table above.

The following table shows a high-level summary score for each tool based on the key capabilities:

TOOL	Data Ingestion and Integration	Data Preparation and Cleaning	Data Profiling, Exploration / Pattern Detection	Data Monitoring	Data Use	TOTAL SCORE
Knime	3.04	4.00	1.00	1.00	1.10	10.14
DataCleaner (COMMUNITY EDITION)	2.79	4.00	1.00	1.00	0.33	9.13
Orange	2.54	3.17	1.00	1.00	0.67	8.38
RapidMiner (LIMITED FREE VERSION)	2.29	1.83	1.00	1.00	0.67	6.79
Aggregate Profiler (AP)	0.29	2.27	0.78	1.00	0.60	4.93
WEKA	0.18	1.72	1.00	0.00	0.60	3.49
Anaconda	0.54	0.33	1.00	1.00	0.50	3.37
Talend Open Studio For Data Quality (LIMITED FREE VERSION)	1.29	1.32	0.56	0.00	0.00	3.17
OpenRefine	1.43	1.05	0.00	0.00	0.00	2.48
SQL Power DQguru	0.29	2.10	0.00	0.00	0.00	2.39
Talend Open Studio For ESB (LIMITED FREE VERSION)	1.29	0.00	0.00	0.00	1.00	2.29
WhiteRabbit	1.34	0.00	0.11	0.33	0.33	2.12
Talend Open Studio for Data Integration (LIMITED FREE VERSION)	2.04	0.00	0.00	0.00	0.00	2.04
Talend Open Studio For Big Data (LIMITED FREE VERSION)	2.04	0.00	0.00	0.00	0.00	2.04
DataMartist (30 DAY FREE TRIAL, STANDARD - \$349, PROFESSIONAL - \$995)	1.29	0.45	0.11	0.00	0.17	2.02
Pandas Profiling (BEST DATA PROFILING)	0.29	0.00	1.00	0.00	0.33	1.63
MobyDQ	0.79	0.00	0.11	0.67	0.00	1.57
Talend Open Studio For MDM (LIMITED FREE EDITION)	0.29	0.98	0.00	0.00	0.17	1.44
DQ Analyzer (COMMUNITY EDITION)	0.29	0.00	1.00	0.00	0.00	1.29
Pentaho Kettle (COMMUNITY EDITION)	1.29	0.00	0.00	0.00	0.00	1.29
DataPreparator	0.43	0.85	0.00	0.00	0.00	1.28
Pimcore	0.00	1.25	0.00	0.00	0.00	1.25
ARX Data Anonymization	0.29	0.58	0.00	0.00	0.33	1.21
CytoScape	0.25	0.00	0.00	0.00	0.64	0.89
Data Match (30-DAY FREE TRIAL)	0.29	0.40	0.00	0.00	0.00	0.69
SQL Power Architect	0.54	0.00	0.00	0.00	0.14	0.69
Anonimatron	0.29	0.00	0.00	0.00	0.17	0.46
pyxplore	0.24	0.00	0.11	0.00	0.00	0.35

The supporting Excel document (**Data Quality Tools Matrix.sls**) can be used to add and score additional tools as needed.

 	Data Ingestion and Integration	Data Ingestion and Integration	Data Ingestion and Integration	Data Ingestion and Integration	Data Preparation and Cleaning	Data Preparation and Cleaning	Data Preparation and Cleaning	Data Preparation and Cleaning	Data Preparation and Cleaning	Data Profiling, Exploration / Pattern Detection	Data Monitoring	Data Use	Data Use
	Connectivity	Parsing	Issue resolution and workflow	Architecture and Integration	Master Reference Data Management	Standardization and cleansing	Matching, linking and merging	Address validation / geocoding	Data curation and enrichment	Data profiling, measurement and visualization	Monitoring	Metadata management	Usability
Knime	0.29	1.00	1.00	0.75	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.43	0.67
Pandas Profiling	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.33
Orange	0.29	1.00	1.00	0.25	0.00	0.50	1.00	1.00	0.67	1.00	1.00	0.00	0.67
RapidMiner	0.29	1.00	0.50	0.50	0.00	0.50	1.00	0.00	0.33	1.00	1.00	0.00	0.67
WEKA	0.18	0.00	0.00	0.00	0.00	0.25	0.80	0.00	0.67	1.00	0.00	0.43	0.17
Anonimatron	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17
ARX Data Anonymization	0.29	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.33	0.00	0.00	0.00	0.33
WhiteRabbit	0.59	0.00	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.11	0.33	0.00	0.33
Aggregate Profiler (AP)	0.29	0.00	0.00	0.00	0.00	0.00	0.60	1.00	0.67	0.78	1.00	0.43	0.17
Talend Open Studio for Data Integration	0.29	1.00	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Talend Open Studio For Big Data	0.29	1.00	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Talend Open Studio For Data Quality	0.29	1.00	0.00	0.00	0.00	0.25	0.40	0.00	0.67	0.56	0.00	0.00	0.00
Talend Open Studio For ESB	0.29	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Talend Open Studio For MDM	0.29	0.00	0.00	0.00	0.40	0.25	0.00	0.00	0.33	0.00	0.00	0.00	0.17
OpenRefine	0.18	1.00	0.00	0.25	0.00	0.25	0.80	0.00	0.00	0.00	0.00	0.00	0.00
DataCleaner	0.29	1.00	1.00	0.50	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.33
DataPreparator	0.18	0.00	0.00	0.25	0.00	0.25	0.60	0.00	0.00	0.00	0.00	0.00	0.00
Data Match	0.29	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00
DataMartist	0.29	1.00	0.00	0.00	0.00	0.25	0.20	0.00	0.00	0.11	0.00	0.00	0.17
Pentaho Kettle	0.29	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SQL Power Architect	0.29	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00
SQL Power DQguru	0.29	0.00	0.00	0.00	0.00	0.50	0.60	1.00	0.00	0.00	0.00	0.00	0.00
DQ Analyzer	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Pimcore	0.00	0.00	0.00	0.00	1.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CytoScape	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.50
Anaconda	0.29	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.33	1.00	1.00	0.00	0.50
pyxplorer	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00
MobyDQ	0.29	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.67	0.00	0.00

Data Quality Tool Matrix

Phase 3: Compare data quality profiling tools

Data processing and analysis can't happen without data profiling, the process of reviewing source data for content and quality. Data profiling is the means to visualize and understand the quality of a dataset. As data gets bigger and infrastructure moves to the cloud, data profiling is increasingly important.

While data profiling is a fundamental step in understanding data quality, it is important to keep in mind that data profiling alone does not create or improve data quality. Specifically, data profiling provides the user with an understanding of the inherent data quality (in various dimensions) within the current data set. Based on the outcome of data profiling, it will likely be required to utilize one or more of the data quality tools evaluated in Phase 2 of this project to improve the quality of the data. This effort is best accomplished by data analysts and/or scientists with subject matter expertise, close to the original source of the data.

1.1 Synthetic Reference Dataset Creation

To evaluate the open data profiling tools, synthetic data sets were created using the open source tool, WhiteRabbit, to generate 1000 patient and related clinical data CSV files and SQL Database adhering to OMOP data model. To evaluate performance and scalability of each tool an additional dataset of 1.3 Million records was also generated using WhiteRabbit.

Additional information about WhiteRabbit, along with installation instruction and sample files can be found in the "WhiteRabbit" directory provided by Inspirata.

1.2 Data Profiling Tools Evaluated

Based on the outcome of the tool summary matrix, the following open data profiling tools were selected to install, run and test against the synthetic data set created with WhiteRabbit.

- KNIME
- DATACLEANER
- ORANGE
- WEKA
- PANDAS-PROFILING (PYTHON)
- AGGREGATE PROFILER
- TALEND OPEN STUDIO FOR DATA QUALITY
- WHITERABBIT (PER REQUEST FROM HDR-UK)

DATA PROFILING TOOLS	TOTAL SCORE
Pandas Profiling (BEST DATA PROFILING)	1.00
Knime	1.00
Orange	1.00
WEKA	1.00
DataCleaner (COMMUNITY EDITION)	1.00
RapidMiner (LIMITED FREE VERSION)	1.00
DQ Analyzer (COMMUNITY EDITION)	1.00
Anaconda	1.00
Aggregate Profiler (AP)	0.78
Talend Open Studio For Data Quality (LIMITED FREE VERSION)	0.56
WhiteRabbit	0.11
DataMartist (30 DAY FREE TRIAL)	0.11
pyxplorer	0.11

In addition, Inspirata evaluated a Data Quality Monitoring/Testing tool called MOBYDQ. MobyDQ is a tool that claims to automate data quality checks on a data ingestion data pipeline, capture data quality issues and trigger alerts in case of anomaly. MobyDQ was extremely hard to install and the results were disappointing.

DATA QUALITY TESTING TOOL	TOTAL SCORE
MobyDQ	0.11

1.3 Data Quality Dimensions

The term data quality dimension has been widely used to describe the measure of the quality of data. In May 2012, DAMA UK assembled a working group to define six Data Quality Dimensions.

<https://www.whitepapers.em360tech.com/wp-content/files/mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf>

Inspirata mapped key data profiling tool capabilities to these data quality dimensions after HDR UK decided that DAMA quality dimensions were appropriate.

- **COMPLETENESS** – The proportion of stored data against the potential of "100% complete". Answers the question: - *"Are all data sets and data items recorded?"*
- **CONSISTENCY** – The absence of difference, when comparing two or more representations of a thing against a definition. Answers the question: - *"Can we match the data set across data stores?"*
- **UNIQUENESS** – Nothing will be recorded more than once based upon how that thing is identified. Answers the question: - *"Is there a single view of the data set?"*
- **VALIDITY** – Data are valid if it conforms to the syntax (format, type, range) of its definition. Answers the question: - *"Does the data match the rules?"*
- **ACCURACY** – The degree to which data correctly describes the "real world" object or event being described. Answers the question: - *"Does the data reflect the data set?"*
- **TIMELINESS** – The degree to which data represent reality from the required point in time. Answers the question: - *"Is data relevant?"*

1.4 Data Profiling Tools Comparison Approach

Evaluating Tool Capabilities

Each of the specified open source data profiling tools were evaluated based on how possible it was to execute the following functions:

Single Column – Cardinalities

REFERS TO THE UNIQUENESS OF DATA VALUES CONTAINED IN A PARTICULAR COLUMN (ATTRIBUTE) OF A TABLE (ENTITY)

FUNCTION	DATA PROFILING TOOLS CAPABLE OF NATIVELY EXECUTING FUNCTION
Number of rows	All evaluated tools
Number of nulls	All evaluated tools
Percentage of nulls	All evaluated tools can provide this information <u>EXCEPT</u> Aggregate Profiler and Data Cleaner
Number of distinct values (cardinality)	All evaluated tools

Percentage of distinct values (Number of distinct values divided by the number of rows)	Knime, WEKA, Pandas Profiling, Talent Open Studio
---	---

Single Column - Value distributions

PRESENTS AN ORDERING OF THE RELATIVE FREQUENCY (COUNT AND PERCENTAGE) OF THE ASSIGNMENT OF DISTINCT VALUES

FUNCTION	DATA PROFILING TOOLS CAPABLE OF NATIVELY EXECUTING FUNCTION
Frequency histograms (equi-width, equi-depth, etc.)	Knime and Pandas Profiling (Python)
Minimum and maximum values in a numeric column	All evaluated tools can provide this information <u>EXCEPT</u> WEKA
Constancy (Frequency of most frequent value divided by number of rows)	Knime, Pandas Profiling, Talent Open Studio
Quartiles (3 points that divide the numeric values into 4 equal groups)	All evaluated tools can provide this information <u>EXCEPT</u> WEKA and Orange
Distribution of first digit in numeric values (to check Benford's law)	Talend Open Studio, Knime, Pandas Profiling (Python)

Single Column - Patterns, datatypes, and domains

REFERS TO THE DISCOVERY OF PATTERNS AND DATA TYPES

FUNCTION	DATA PROFILING TOOLS CAPABLE OF NATIVELY EXECUTING FUNCTION
Basic types (e.g., numeric, alphanumeric, date, time)	Knime and Pandas Profiling
DBMS-specific data type (e.g., varchar, timestamp)	All evaluated tools can provide this information <u>EXCEPT</u> Orange and WEKA
Measurement of Value length (minimum, maximum, average, median)	All evaluated tools can provide this information <u>EXCEPT</u> Talend Open Studio and WEKA
Maximum number of digits in numeric values	Aggregate Profiler, Data Cleaner, Knime, Pandas (Python)
Maximum number of decimals in numeric values	Aggregate Profiler, Knime, Pandas (Python)

Histogram of value patterns (Aa9...)	Data Cleaner, Talend Open Studio, Knime, Pandas (Python)
Generic semantic data type (e.g., code, date/time, quantity, identifier)	Data Cleaner, Talend Open Studio, Knime, Pandas (Python)
Semantic domain (e.g., credit card, first name, city)	Data Cleaner, Talend Open Studio, Knime, Pandas (Python)

Dependencies

DETERMINES THE DEPENDENT RELATIONSHIPS WITHIN A DATA SET

FUNCTION	DATA PROFILING TOOLS CAPABLE OF NATIVELY EXECUTING FUNCTION
Unique column combinations (UCCs) (key discovery)	Pandas Profiling (Python)
Relaxed unique column combinations	Pandas Profiling (Python)
Inclusion dependencies (INDs) (foreign key discovery)	Pandas Profiling (Python)
Relaxed inclusion dependencies	Pandas Profiling (Python)
Functional dependencies	Pandas Profiling (Python)
Conditional functional dependencies	Pandas Profiling (Python)

Advanced Multi Column profiling

DETERMINES THE SIMILARITIES AND DIFFERENCES IN SYNTAX AND DATA TYPES BETWEEN TABLES (ENTITIES) TO DETERMINE WHICH DATA MIGHT BE REDUNDANT AND WHICH COULD BE MAPPED TOGETHER

FUNCTION	DATA PROFILING TOOLS CAPABLE OF NATIVELY EXECUTING FUNCTION
Correlation analysis	Pandas Profiling (Python), Aggregate Profiler and Orange
Association rule mining	Pandas Profiling (Python)
Cluster analysis	Pandas Profiling (Python)
Outlier detection	Pandas Profiling (Python), Orange and Knime
Exact duplicate tuple detection	Pandas Profiling (Python), Data Cleaner and Talend Open Studio
Relaxed duplicate tuple detection	Pandas Profiling (Python), Data Cleaner and Talend Open Studio

Scoring Results

Subsequently, each tool was evaluated in terms of how well it delivers each key capability required to address data quality on a five-point scale.

0 = Not applicable
1 = Poor: most or all defined requirements not achieved
2 = Fair: some requirements not achieved
3 = Good: meets requirements
4 = Excellent: meets or exceeds some requirements
5 = Outstanding: significantly exceeds requirements

■ COMPLETENESS

The proportion of stored data against the potential of "100% complete"

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
Percentage of requisite information available	2	4	4	3	2	3	5	1	1
Percentage of missing data values (null / empty string)	2	4	4	4	3	3	5	1	0
Row counts	4	5	4	4	4	3	5	2	1
Highest and lowest value of key elements	0	3	5	0	0	3	5	1	0
Number of data values in an unusable state	0	2	2	0	0	3	5	0	0

■ UNIQUENESS

No thing will be recorded more than once based upon how that thing is identified.

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
(Number of things in the real world) - Number of incorrect spellings etc. of same data in an element e.g. address (duplicate values)	0	2	2	0	1	2	5	2	0
(Number of recodes describing different things) Number of data items in adherence to expected/described data element value (distinct values at ID level)	0	1	2	0	1	2	5	1	0
(Number of things in real world i.e. duplicates)/(Number of records describing different things i.e. distinct records)	0	3	4	4	1	2	5	1	0

■ TIMELINESS

The degree to which data represent reality from the required point in time.

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
Difference between Lowest date value and Highest Date Value	0	2	4	0	1	2	3	1	1
Number of records per month	0	1	3	0	0	2	3	0	0

■ VALIDITY

Data are valid if it conforms to the syntax (format, type, range) of its definition.

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
Percentage of data values that comply with the specified formats (data types, ranges etc)	0	1	3	0	0	4	5	2	0
Percentage of data values that don't comply to specified formats	0	0	1	0	0	1	4	0	0
Number of Missing values indicated e.g. with fill values	0	4	4	0	4	3	5	2	0
Number of Values in Specified Range	0	0	3	0	0	3	4	0	0
Number of values not in Specified Range	0	0	2	0	0	3	3	0	0

■ ACCURACY

The degree to which data correctly describes the "real world" object or event being described.

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
Number of accurate data values	0	3	3	0	2	0	5	2	0
Number of inaccurate data values	0	0	0	0	0	0	5	0	0
Actual data value count versus predicted data value count	0	0	0	0	0	0	3	0	0
Number of rows and columns against expectations	0	0	0	0	0	0	3	0	0
Number of duplicates at ID level	0	4	4	4	3	3	5	3	0
Number of blank columns, large % of blank data, high % of same data	0	3	4	0	2	0	5	2	0
Distribution across various segments	0	3	0	0	0	0	5	0	0
Outliers on key variables	0	3	2	0	0	0	4	0	0
((Count of accurate objects)/ (Count of accurate objects + Counts of inaccurate objects))	0	1	1	0	0	0	3	0	0

■ CONSISTENCY

The absence of difference, when comparing two or more representations of a thing against a definition.

Measure (key elements)	White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
Analysis of pattern and/or value frequency	0	0	0	0	0	0	5	0	0

1.5 Data Profiling Tools Comparison Summary Results

The supporting Excel document (***ProfilingToolsToQualityDimensionMatrix.xlsx***) contains details form data profiling outcome.

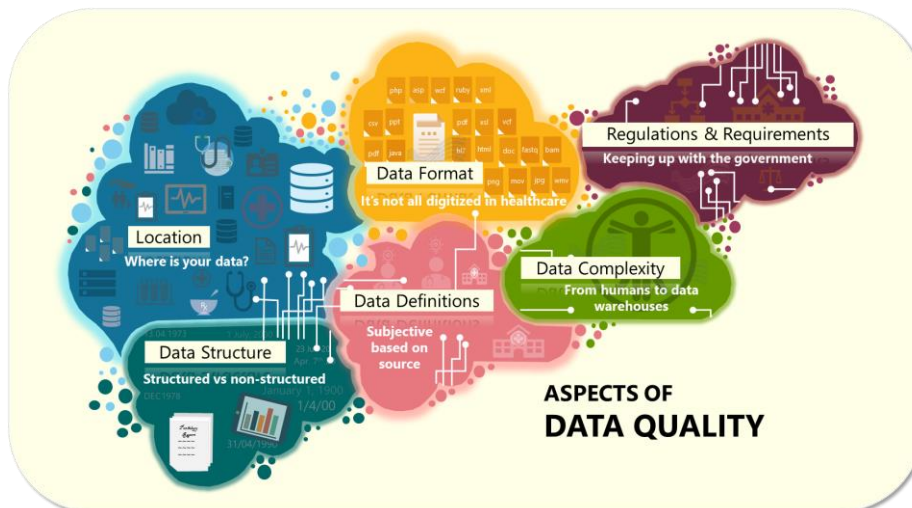
Inspirata found the most comprehensive and scalable solutions to be a combination:

- Knime augmented with Pandas-Profiling **OR**
- Orange (MacOS, Linux) augmented with Pandas Profiling

White Rabbit	Orange	Knime	WEKA	Aggregate Profiler	DataCleaner	Pandas (Python)	Talend Open Studio - Data Quality	MobyDQ
2	4	4	3	2	1	5	1	1
2	4	4	4	3	1	5	1	0
4	5	4	4	4	1	5	2	1
0	3	5	0	0	1	5	1	0
0	2	2	0	0	1	5	0	0
0	2	2	0	1	1	5	2	0
0	1	2	0	1	1	5	1	0
0	3	4	4	1	1	5	1	0
0	2	4	0	1	1	3	1	1
0	1	3	0	0	1	3	0	0
0	1	3	0	0	1	5	2	0
0	0	1	0	0	1	4	0	0
0	4	4	0	4	1	5	2	0
0	0	3	0	0	1	4	0	0
0	0	2	0	0	1	3	0	0
0	3	3	0	2	1	5	2	0
0	0	0	0	0	1	5	0	0
0	0	0	0	0	1	3	0	0
0	0	0	0	0	1	3	0	0
0	4	4	4	3	1	5	3	0
0	3	4	0	2	0	5	2	0
0	3	0	0	0	0	5	0	0
0	3	2	0	0	0	4	0	0
0	1	1	0	0	1	3	0	0
0	0	0	0	0	1	5	0	0

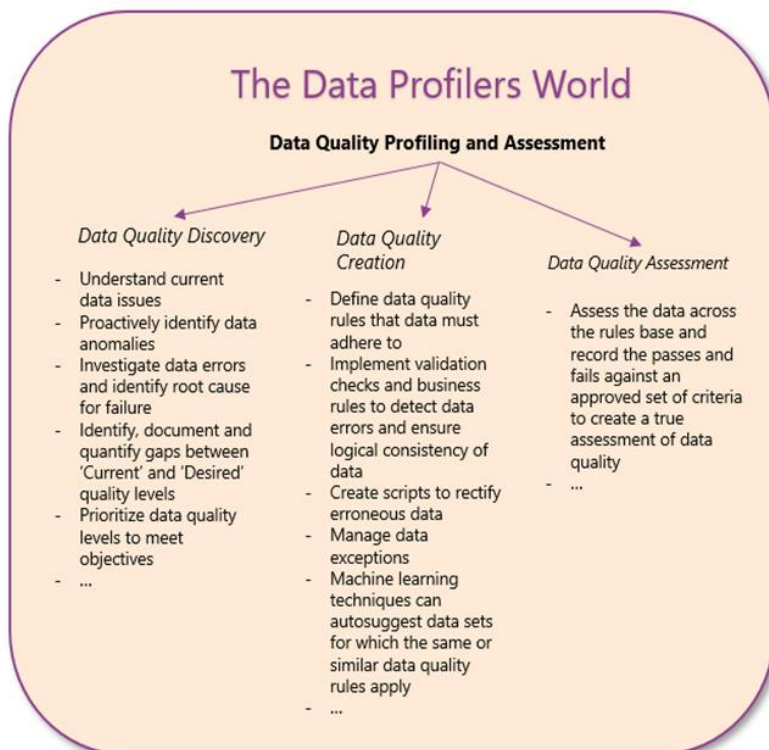
1.6 The Data Quality Process

Healthcare data is challenged by many different aspects of data quality.



Subject matter experts are essential to ensuring data quality is achieved as data moves from a Raw-to-Staged-to-Trusted-to-Published state.

1.7 Skills needed for Data Profiling



Data Profiling is only one aspect of ensuring high data quality. It is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects. The results from data profiling help organisations to make important decisions by identifying various facts and trends.

The Data Profiler

Answers:

- What is required?

Actions:

- Profiles, analyzes and inspects data to identify patterns, augments and enhances data, metadata and schemas.
- Documents data element metadata, data flow and related relationships.
- Provides scripted business rules and algorithms to augment or analyze data.
- Does "Endogenous data enrichment" by transforming existing data into derived variables that are more informative and meaningful relative to the questions being asked than the original data.

Expertise:

Typically, this persona is a **data scientist** with expertise in the specific healthcare field subject matter. Intermediate to advanced technical skills, ability to create scripts using languages like Python and R, well versed in ML.

Results:

Defines **meaningful data**. Documents how cleansed data accomplishes business needs and identifies and documents any data quality gaps.

Data profiling is a crucial part of:

- **Data warehouse and business intelligence (DW/BI) projects**—data profiling can uncover data quality issues in data sources, and what needs to be corrected in ETL.
- **Data conversion and migration projects**—data profiling can identify data quality issues, which you can handle in scripts and data integration tools copying data from source to target. It can also uncover new requirements for the target system.
- **Source system data quality projects**—data profiling can highlight data which suffers from serious or numerous quality issues, and the source of the issues (e.g. user inputs, errors in interfaces, data corruption)

It is important that the data analyst responsible for executing data profiling be skilled in data profiling techniques and associated tools.

Key skills for a data analyst:

- A high level of mathematical ability
- Programming languages, such as SQL, Oracle and Python
- The ability to analyse, model and interpret data
- Problem-solving skills
- A methodical and logical approach
- The ability to plan work and meet deadlines
- Accuracy and attention to detail
- Interpersonal skills
- Teamworking skills
- Written and verbal communication skills

Open Data Quality Profiling Tools Overview, Features, Strengths & Limitations

Each of the Data Quality Profiling tools was installed and tested against the synthetic dataset generated with WhiteRabbit.

ADDITIONAL INFORMATION RELATED TO DATA PROFILING OPEN TOOLS REGARDING HOW TO INSTALL AS WELL AS DIFFERENT FINDINGS AND SCRIPTS WHERE APPROPRIATE CAN BE FOUND FOR EACH TOOL IN THE APPROPRIATE DIRECTORY AS PROVIDED BY INSPIRATA.

- 1.1 Knime
- 1.2 Orange
- 1.3 Pandas Profiling (Python)
- 1.4 WEKA
- 1.5 Aggregate Profiler
- 1.6 DataCleaner
- 1.7 Talend Open Studio for Data Quality
- 1.8 MobyDQ



During the project, the Cystic Fibrosis Trust and Neonatal Medicine Research Group, Imperial College London teams agreed to test two data profiling tools each and time permitting, MobyDQ. Inspirata would sincerely like to thank the following folks for their willingness to evaluate and run the specified tools and for providing their feedback to HDR-UK and Inspirata.

This feedback has been incorporated into the above-mentioned supporting documents where applicable

- Cystic Fibrosis Trust tested "Aggregate Profiler", "Knime" and "MobyDQ":
 - **Kieran Earlam**, Policy and Evidence Manager
 - **Rebecca Cosgriff**, Director of Data & Quality Improvement
- Neonatal Medicine Research Group tested "DataCleaner", "Orange" and "MobyDQ":
 - **Victor L Banda**, Data Analyst from the Neonatal Data Analysis Unit

Info Cards - Other Open Source Data Quality Tools

The supporting Excel document (***Data Quality Tools Matrix.sls***) can be used to view additional information about all the tools compared, by clicking on the tool name or within the same row:

	Data Ingestion and Integration	Data Ingestion and	Data Ingestion and	Data Ingestion and	Data Preparation	Data Preparation	Data Preparation	Data Preparation	Data Preparation	Data Profiling, Exploration / Pattern
<div>  Add Tool  Delete Tool </div>	Info Card									
	Connectivity									
Knime	0.29									
Pandas Profiling	0.29									
Orange	0.29									
RapidMiner	0.29									
WEKA	0.18									
Anonimatron	0.29									
ARX Data Anonymization	0.29									
WhiteRabbit	0.59									
Aggregate Profiler (AP)	0.29									
Talend Open Studio for Data Integration	0.29									
Talend Open Studio For Big Data	0.29									
Talend Open Studio For Data Quality	0.29									
Talend Open Studio For ESB	0.29									
Talend Open Studio For MDM	0.29									
OpenRefine	0.24									
DataCleaner	0.29									
DataPreparator	0.24									

Data Quality Tool Matrix

WEKA

<https://sourceforge.net/projects/weka/>
<https://www.cs.waikato.ac.nz/ml/weka/>

License

GNU General Public License version 3.0 (GPLv3)

Version

3.8.4

Last Update

System Requirements

Supported OS

Windows, macOS, Linux

Description

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can be used via a graphical user interface, standard terminal applications, or a Java API. It is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute type using Java Database Connectivity and can process the result returned by a database query).

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute type using Java Database Connectivity and can process the result returned by a database query).

Weka is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table suitable for processing covered by the algorithms included in the Weka distribution is sequence modeling.

FEATURES:

- machine learning
- data mining
- preprocessing
- classification

Selected Features

Select Features

Conclusion

There are two critical aspects of data that must be understood: the meaning and the quality. Deficiencies in either can adversely impact the data's capacity to support research conclusions and in many cases may render data altogether unusable for downstream use cases or studies.

Data are always an incomplete representation of the things and events they describe, and as such may be appropriate for some uses but inadequate for others. Data quality is a key component of data utility. Data in a given data set may have an acceptable level of quality in some contexts but be inadequate in other contexts.

The further removed the analysis team is from the original data source and collection (date or process), the greater potential there is for misunderstanding, degradation, loss of information, or misuse. For example, medical data that have been dictated by a clinician in a discharge letter and subsequently coded with a standard terminology (e.g. International Classification of Diseases [ICD], SNOMED or Current Procedural Terminology [CPT]) represent processed data that are removed from their origin. Such data are at risk for information loss through data reduction from coding, through disassociation with contextual information, or through the introduction of errors. Thus, while users of secondary data may not have control over the original data collection, it is important they understand how and why those data were originally obtained, as well as any subsequent processing to which they were subjected.

Sound decisions are based on sound data and it is therefore essential to have Subject Matter Expert involved in the assessment and analysis to ensure:

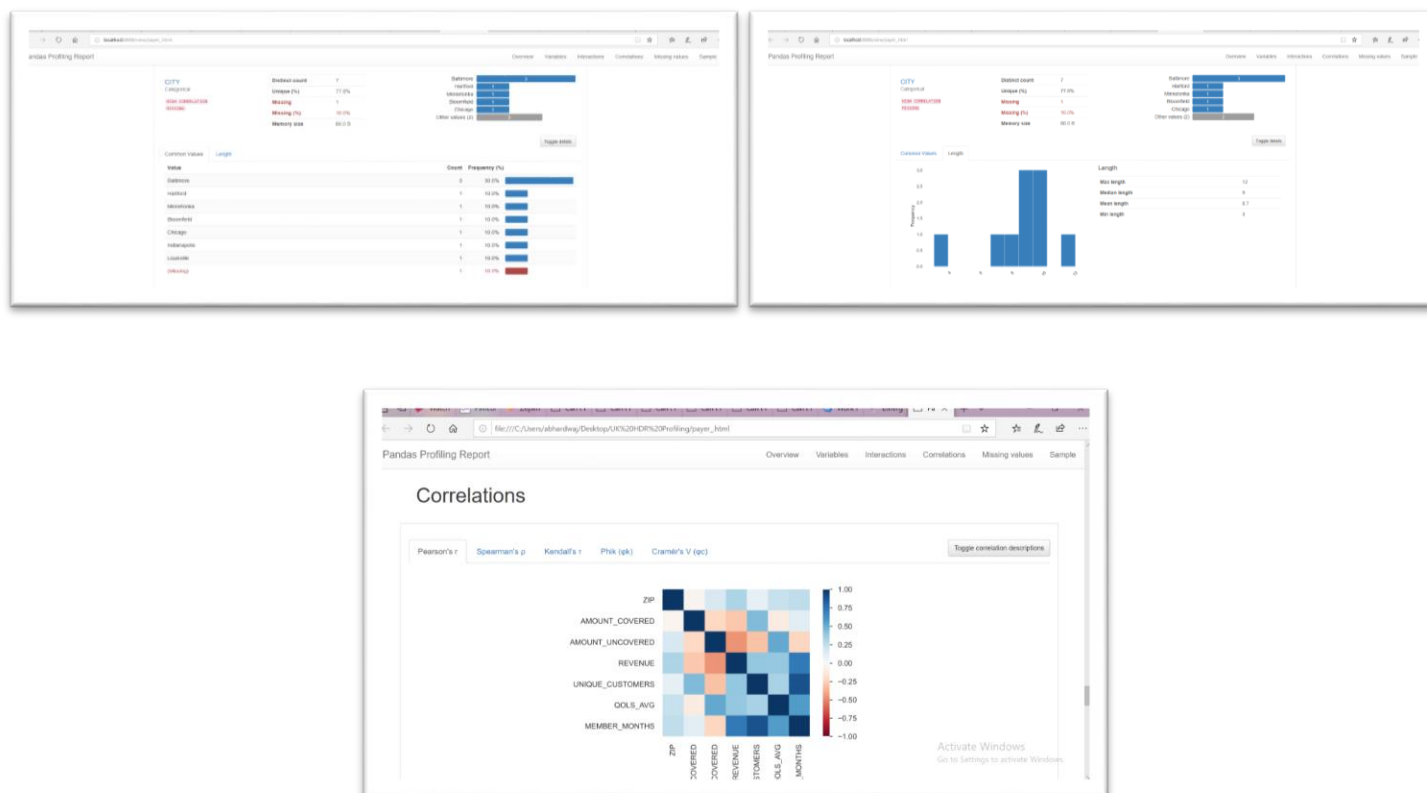
- Data is complete and timely – i.e. there is sufficient information available to make informed and accurate decisions.
- Data is consistent and reliable – i.e. data are plausible and consistent on repeated measurement.
- Data is accurate – i.e. data is medically accurate and hasn't been wrongfully altered in any way.

Based on Inspirata's extensive experience in the healthcare data field, we highly recommend that the data analyst(s) responsible for executing data profiling are domain-specific experts, skilled in data profiling techniques and associated tools.

Inspirata assessed various open data profiling tools and found the most comprehensive and scalable solutions to be a combination:

Knime augmented with Pandas-Profiling
OR
Orange augmented with Pandas Profiling

The following depicts but a few of the excellent visualizations produced by Pandas Profiling. Pandas Profiling produces a .HTML version of these reports which could be easily integrated onto the Innovation Gateway.



Recommendations

HDR UK's mission is focused on **uniting, improving, and using** health data to improve people's lives. HDR UK has established eight Health Data Research Hubs, that along with 30 organizations across the UK Health Data Research Alliance, have published 442 datasets on the Health Data Research Innovation Gateway.

During the four months span of this project, we observed and quantified a significant improvement in the quality of metadata associated with the datasets on the Innovation Gateway. This improvement directly resulted from having a baseline of the quality of the initial metadata, and systematically working across the Alliance to make improvements to the metadata. The second assessment of metadata quality we performed, while showing a huge improvement, also shows that there is room for further improvements. **It is recommended to maintain the focus on improving metadata quality and to assess those improvements using the approach adopted herein on an on-going, periodic basis.**

Beyond the metadata, this project has laid the foundation for assessing data quality through data profiling tools, as well as making incremental improvements in data quality by adopting robust processes and data quality [creation] tools at the sources of the data. **We recommend to conduct data profiling of all 442 datasets on the Innovation Gateway, as well as any other datasets at Hubs or with data custodians that have not yet been published to the Gateway.** By doing so, HDR will establish a baseline of data quality for all datasets against which improvements in data quality can be measured. Additionally, the

output from data profiling tools may be useful in providing consumers of data, accessing datasets via the Innovation Gateway, with meaningful insights to the dataset they are interested in. As such, **our recommendation is to explore which data profiling reports and visual representations thereof will be most useful to data consumers and determine means to integrate those onto the Innovation Gateway.**

Another finding, during this project has been that some of the data custodians, while they have perhaps the responsibility to collect and manage specific datasets, they do not have the data engineering expertise at hand to successfully perform data profiling operations. This finding is exemplified by the experience of the Cystic Fibrosis team who diligently performed the assessment of the tooling recommended by Inspirata but admitted that performing such tasks was not within their skillset, nor the skillset of many other data custodians. With this in mind, **our recommendation is to work with the data custodians to provide them with scripted workflows for their specific datasets, to enable custodians who lack data engineering expertise to reasonably profile their data.**

Finally, as mentioned earlier in this report, data profiling is a necessary, but insufficient step alone towards data quality. With an understanding of the deficiencies inherent within a dataset, data engineers typically build and/or adjust procedures and tools to optimize the quality of the dataset they are collecting, curating and/or managing. This project has provided the basis for selecting appropriate tools for creating and improving data quality, although the specific choices will be data, environment and custodian specific. Given that many data custodians lack deep data engineering expertise, **it is recommended to engage an organization that can work in a highly collaborative fashion with data custodians and subject matter experts to create robust data quality processes.**

It has been Inspirata's privilege to work with and serve HDR-UK and we hope to continue working with you to improve UK health data and by doing so positively impact people's lives.