castor

# The Role of Structured Data Science in DCT



castor

# Table of Contents


4 **The role of structured data**

5 **The role of unstructured data**

6 **The role of FAIR data principles**

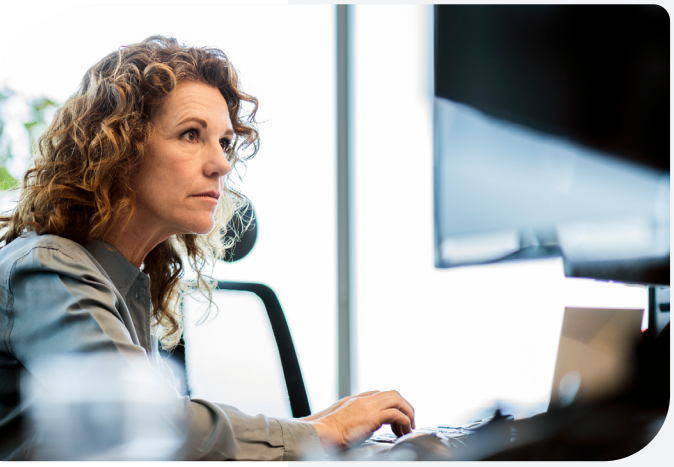
7 **The role of data transformation**




*As clinical research moves forward, there's increased focus on the need for findable, accessible, interoperable, and reusable data. In this article, we'll look at the roles of structured, unstructured data, FAIR data principles, and data transformation in today's data science.*

Data collection and analysis are foundational to advancing medicine—but a shocking 80% of research data isn't reusable. Although today's technologies allow for medical data to be collected and stored digitally, there are obstacles to deriving the maximum value possible from this hard-won resource. One of the main reasons is that most data sets are not machine-actionable or machine-readable.

Early EDC systems were built on paper-based processes and were not designed to collect unstructured health data, such as images and bulk data from remote wearables. Also, attempts to make data findable, accessible, interoperable, and reusable often happen after the data are collected. In this article, we'll dig into the role of structured and unstructured data. We'll also look at how the FAIR data principles can solve some of the nagging issues in data science.



## Machine-actionable vs machine-readable

Data that is machine-actionable can be resolved by web services. In contrast, machine-readable data is in a format automatically read and processed by a computer. The process of making data more machine-readable and machine-actionable continues to be a major obstacle for data management personnel and researchers.

# The role of structured data

Structured data is ordered, predictable, and fits neatly into a specific framework. For example, patient weight is a continuous variable within a predefined range (e.g. 50 to 300) with pounds or kilos dictated. Structured data can neatly slot into a spreadsheet with column names and rows without any extra work. In clinical trials, structured data sits in a file structured by variables such as patient name, gender, weight, medication, blood pressure, etc. Analysis tools incorporating machine learning can easily sort through this kind of organized data and draw accurate conclusions by comparing categorical variables.

## Understanding Relational Database Management Systems (RDBMS)

A relational database management system (RDBMS) is a database management system (DBMS) based on the relational model introduced by E. F. Codd. RDBMS is the basis for SQL and all modern database systems like MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.

### SQL databases

Structured Query Language (SQL) is a language that operates databases and includes database creation, deletion, fetching rows, and modifying rows. SQL is popular because it allows users to:

- Access data in the relational database management systems
- Describe the data
- Define the data in a database and manipulate that data
- Create and drop databases and tables
- Create view stored procedures and functions in a database
- Set permissions on tables, procedures, and views



# The role of unstructured data

Unstructured data is information that is not predefined through data models. For example, it could be the content found inside an x-ray or a voice recording. Unlike its structured counterpart, unstructured data is not yet categorized or tagged. Consequently, when sifting through unstructured data AI (and even human researchers) cannot immediately decipher if it is comparing apples to apples, apples to oranges—or if they're discussing fruit at all.

Consider one example of the challenges associated with unstructured data in a clinical trial. Clinical trials may use information from a patient's electronic medical record (EMR). But since hospital charts are not typically set up with the end goal of clinical research, they often include unstructured data, such as freeform doctors' notes. For instance, if a patient has a tumor, a physician may dictate a note reporting its size and location. But since physicians may use a variety of medical terminology or even size metrics, it can be challenging to identify and capture consistent tumor data if this data is ever analyzed as part of a study.



Storing data in an unstructured way—with the intent to structure it later—might work for a small amount of data, but it can quickly become unwieldy as the amount of data grows. As participants use remote wearables and other mobile health solutions, they produce an unprecedented amount of unstructured data each day. Researchers need a way of organizing, or structuring, their clinical trial data.

It is almost impossible, however, to create a general-purpose algorithm when dealing with unstructured data. For example, how can a program make sense of the information provided by an accelerometer? What counts as a step versus another kind of movement? This leads to the need for annotations, which is the process of labeling data in various formats (e.g., text, video, and images). It can include drawing lines around tumors, pinpointing cells, or interpreting ECG rhythm strips. In some cases, this means paying medical doctors to read thousands of medical records to codify diagnoses. This info can then train an AI system to learn what an image represents via shape analysis, gradients, etc. and compare it to the given diagnosis. For example, melanoma studies can use AI applications to sort through images and accurately flag the ones representing melanoma. The goal is to give an AI system enough correctly-identified training material that it can accurately identify and categorize unstructured data.

# The role of FAIR data principles

Given the enormous amount of unstructured data generated each day as part of clinical trials, the task of structuring it for analysis might seem overwhelming. Fortunately, there's a better way forward. As clinical researchers navigate through a growing amount of unstructured data in modern trials, they may find the FAIR principles helpful. In 2016, a diverse group of researchers published the FAIR Data Principles stating that research data should be:



**Findable**



**Accessible**



**Interoperable**



**Reusable**

These principles promote efficient analysis of data, improve the reliability of the data collection, and allow the data to be reused in future trials. (Learn more about how ongoing informed consent is involved here.) They apply to both computers (machine-actionable and machine-readable) and humans.

### Requirements for FAIRification via an EDC system:

1. The method should be platform-agnostic.
2. The method should not require additional technical knowledge from end-users.
3. eCRF data should be transformable into a machine-readable format.
4. The transformed data must be able to be exposed and queried.
5. The data and metadata must be exposed in a manner following the FAIR Data Principles.



Recently, Castor embarked on a trial project to explore how eCRF data entered into its EDC can be transformed into machine-readable, FAIR data using a semantic data model and mappings from the model to questions on the eCRF. A research team took information from a rare disease registry and successfully applied a FAIRification method in Castor's EDC system. The method mainly focused on data transformation steps to make eCRF data actionable and readable by machines.

It's going to take a while before everyone makes their data FAIR at the source. As Castor demonstrated in the above example, there are possibilities for FAIRification—a necessary capability in a time where not all systems can generate FAIR data.

# The role of data transformation

Currently, efforts to make data (more) machine-readable and FAIR are usually carried out after the research project is conducted and data are collected. The good news is that it's possible to set up FAIRification workflows that can be used from the beginning of a project by researchers who are not data scientists. During the project mentioned above, data entered on eCRFs was automatically transformed into machine-readable data using a FAIRification tool and data model describing the relationships between every collected data element. This meant that the data could be accessed and queried using SPARQL queries. To do so:

1. The FAIRification tool **retrieved records (patient cases)** using the EDC system's API.
2. The FAIRification tool then got **the eCRF data** of every record using the same API.
3. The FAIRification tool then used the  data model to generate a patient-specific graph, where all the relationships between collected data for this patient are machine readable (e.g., Patient 1 → has biological sex → Male).

In Castor's FAIRification tooling, data models do not need to be hardcoded. Instead, they can be added in a user-friendly interface and then mapped to questions on the eCRF by the person that is building the eCRF. Best of all, this does not require any technical knowledge. And after the data model is entered in the data transformation application and mapped to the eCRF, the application does not need any intervention from data entry or data management personnel.

As research becomes more data-intensive and payers focus increasingly on real-world evidence, the role of unstructured data—and the need for it to be FAIR—will only rise. Currently, the legal and regulatory aspects of unstructured data are complex and differ by country.

**Want to know more about how Castor's API can help you meet the challenges of unstructured data? Contact us, or schedule a demo today!**

▶ castoredc.com



