Towards a Big Data Taxonomy

Bill Mandrick, PhD

Data Tactics

Version 26_August_2013

Scientific Taxonomies Represent

- Types of Processes
- Types of Objects
  - Physical Objects
  - Information Artifacts
- Types of Characteristics
  - Qualities
  - Roles
- Relationships
  - Between Processes
  - Between Objects
  - Between Characteristics

Big Data Taxonomy

- Big Data Related Processes
- Big Data Characteristics
- Big Data Information Artifacts
- Big Data Information Bearers
- Relationships between Big Data Elements
- Mapping Instances to the Taxonomy
- Creating Situational Awareness

Relations Between Processes

- Processes A <relation> Processes B
  - Complex Process <has part> Sub-Process
  - Sub-Process <part of> Complex Process
  - Process A <precedes> Process B
  - Process A <follows> Process B

Examples:

Data Curation Process <has part> Data Selection Process
Data Curation Process <has part> Data Collection Process
Data Curation Process <has part> Data Archiving Process

Information Artifact Lifecycle Processes

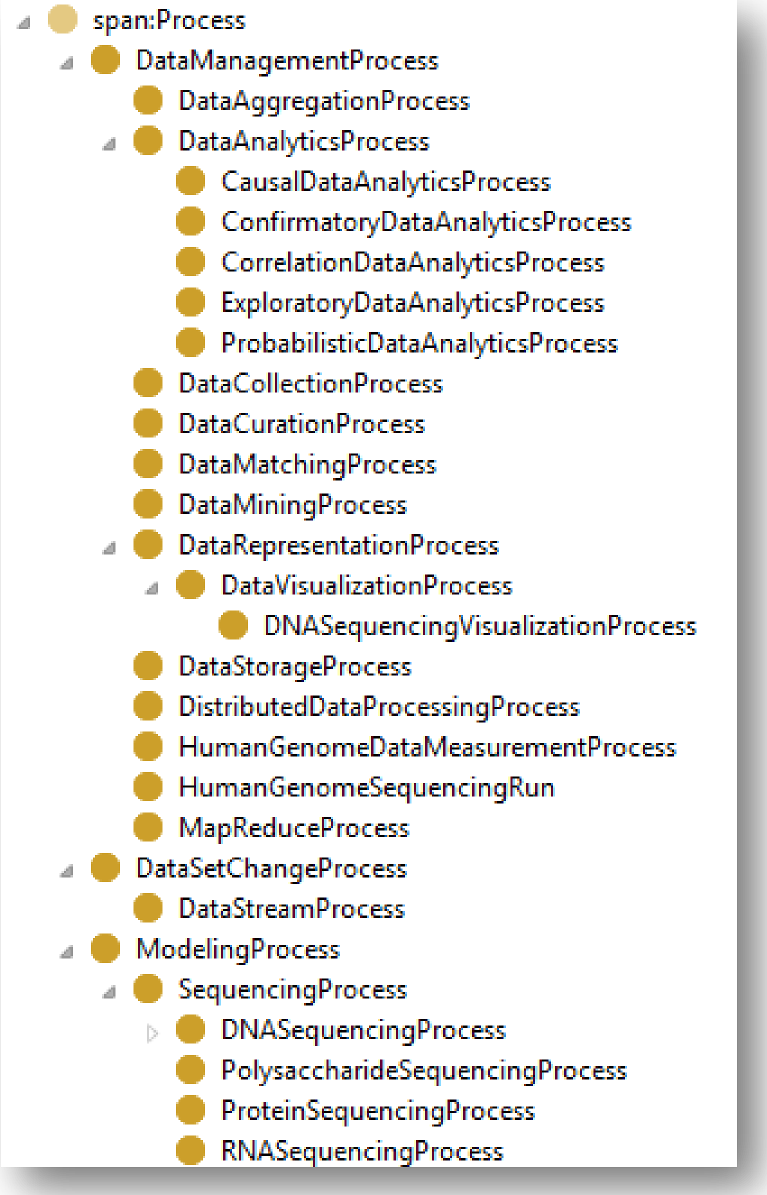
Common Labels

- Collecting
- Curating
- Representing
- Storing
  - Cluster Storing
- Managing
  - Processing
    - Distributed Processing
      - Map Reduce
- Analyzing
  - Data Mining
  - Causal Analysis
  - Probabilistic Analysis
  - Correlation Analysis

Taxonomy Labels

- Data Collection Process
- Data Curation Process
- Data Representation Process
- Data Storing Process
  - Cluster Storing Process
- Data Management Process
  - Processing
    - Distributed Data Process
      - Map Reduce Process
- Data Analytics Process
  - Data Mining Process
  - Causal Analysis Process
  - Probabilistic Analysis Process
  - Correlation Analysis Process

Big Data Processes

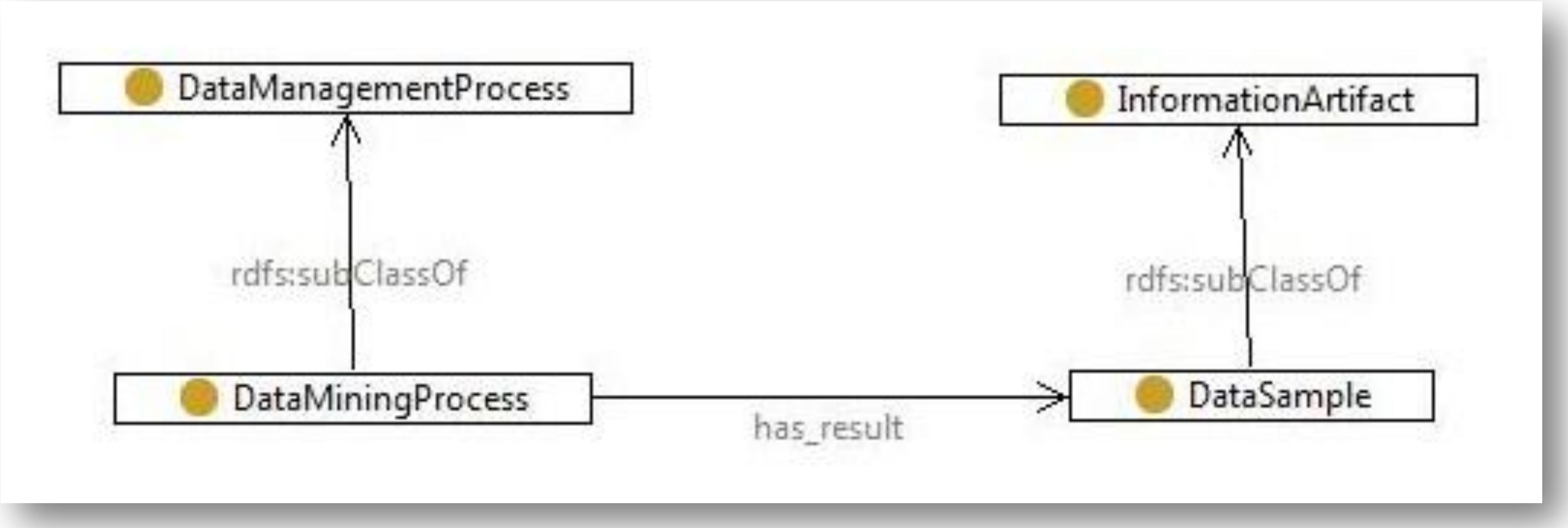


- span:Process
  - DataManagementProcess
    - DataAggregationProcess
    - DataAnalyticsProcess
      - CausalDataAnalyticsProcess
      - ConfirmatoryDataAnalyticsProcess
      - CorrelationDataAnalyticsProcess
      - ExploratoryDataAnalyticsProcess
      - ProbabilisticDataAnalyticsProcess
    - DataCollectionProcess
    - DataCurationProcess
    - DataMatchingProcess
    - DataMiningProcess
    - DataRepresentationProcess
      - DataVisualizationProcess
        - DNASequencingVisualizationProcess
    - DataStorageProcess
    - DistributedDataProcessingProcess
    - HumanGenomeDataMeasurementProcess
    - HumanGenomeSequencingRun
    - MapReduceProcess
  - DataSetChangeProcess
    - DataStreamProcess
  - ModelingProcess
    - SequencingProcess
      - DNASequencingProcess
      - PolysaccharideSequencingProcess
      - ProteinSequencingProcess
      - RNASequencingProcess

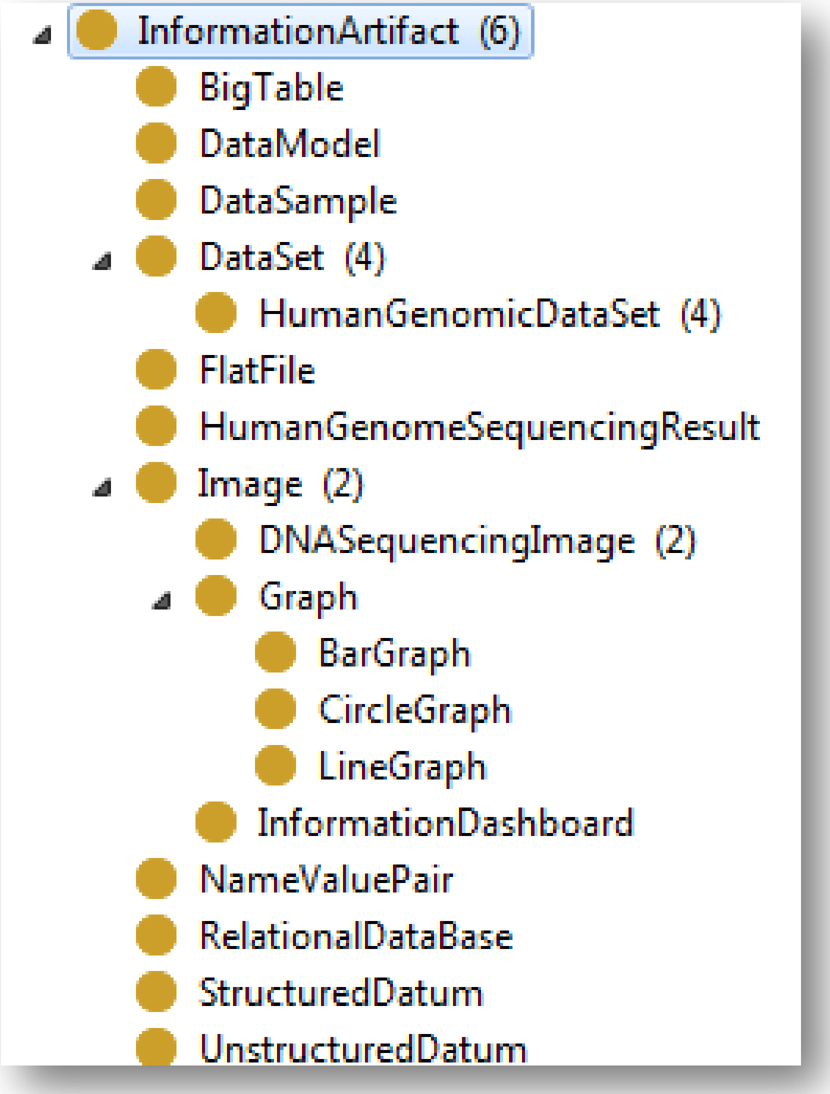
Big Data Processes can be decomposed and related to other (sub)processes

…as well as to their outputs (Information Artifacts).

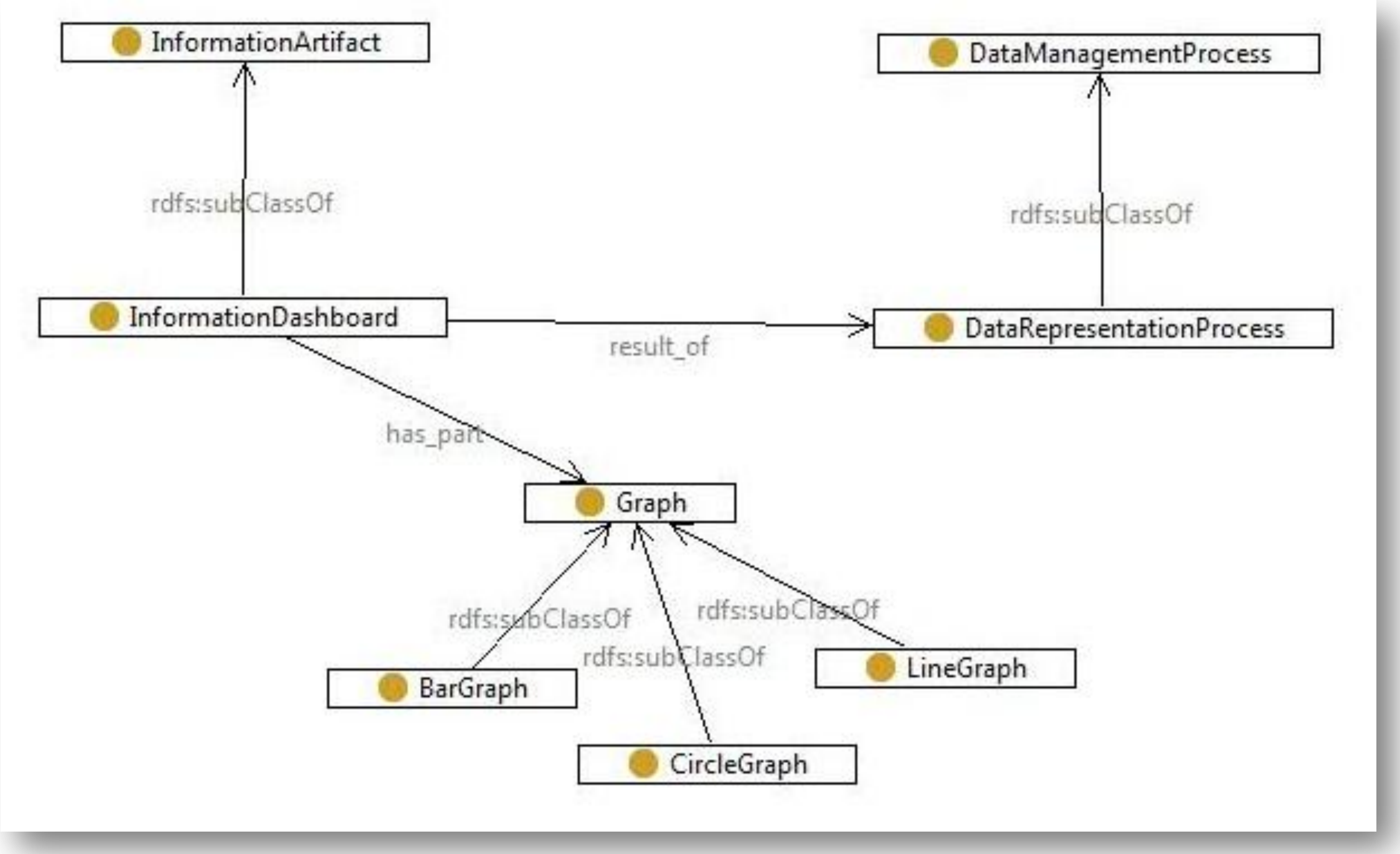
Relating Processes to Products

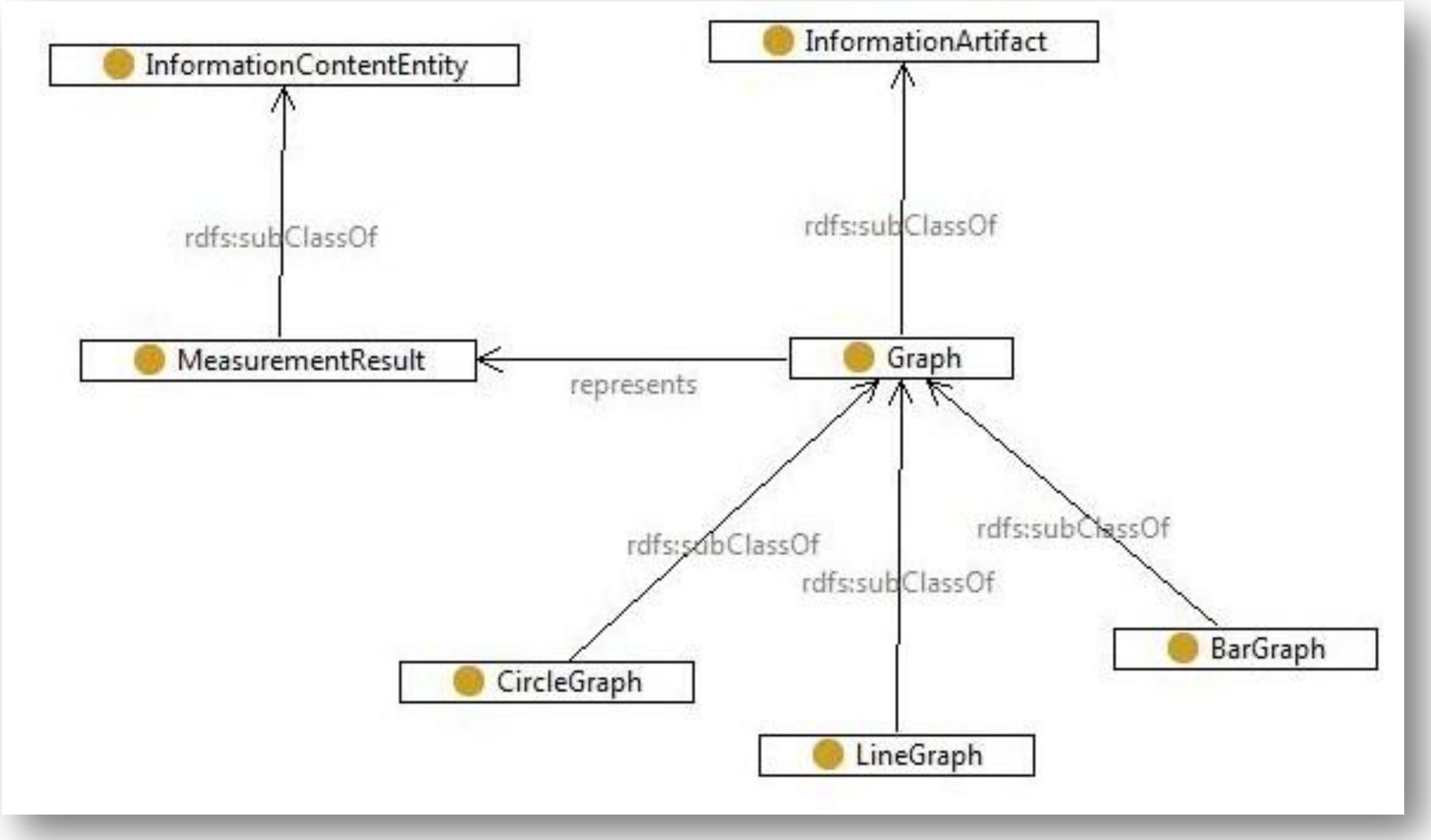


Big Data Information Artifacts

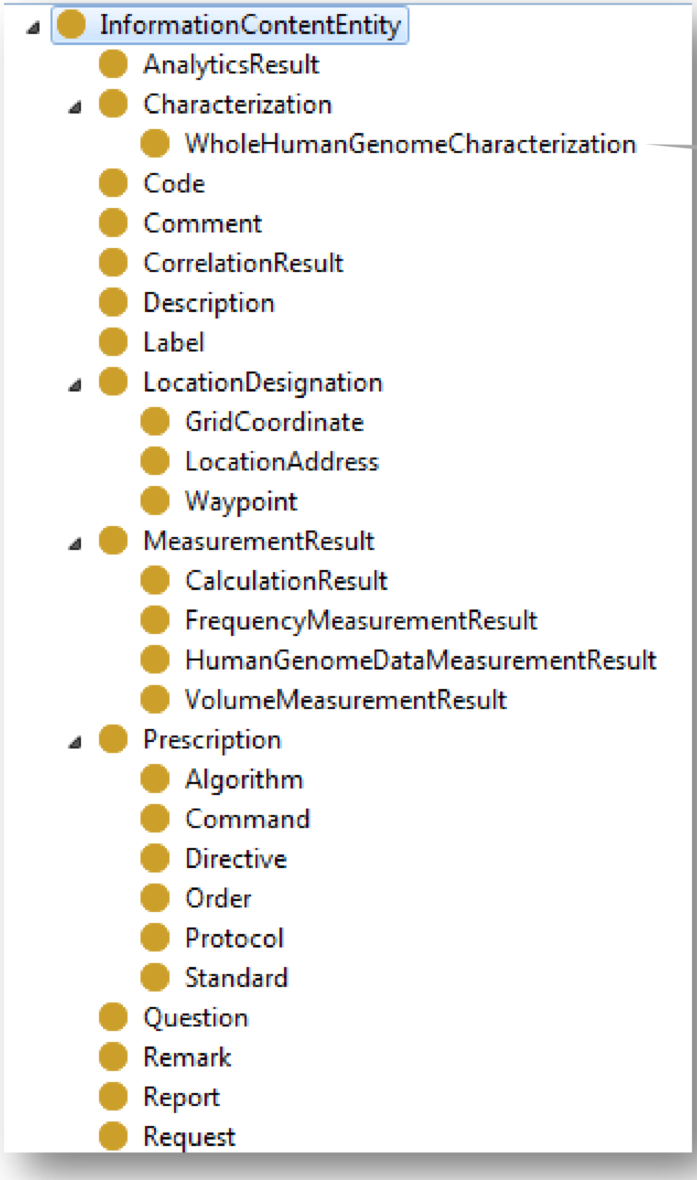


- InformationArtifact (6)
  - BigTable
  - DataModel
  - DataSample
  - DataSet (4)
    - HumanGenomicDataSet (4)
  - FlatFile
  - HumanGenomeSequencingResult
  - Image (2)
    - DNASequencingImage (2)
    - Graph
      - BarGraph
      - CircleGraph
      - LineGraph
    - InformationDashboard
  - NameValuePair
  - RelationalDataBase
  - StructuredDatum
  - UnstructuredDatum


InformationArtifact
DataManagementProcess
rdfs:subClassOf
rdfs:subClassOf
InformationDashboard
result_of
DataRepresentationProcess
has_part
Graph
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
BarGraph
LineGraph
CircleGraph


InformationContentEntity
InformationArtifact
rdfs:subClassOf
rdfs:subClassOf
MeasurementResult
Graph
represents
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
BarGraph
CircleGraph
LineGraph

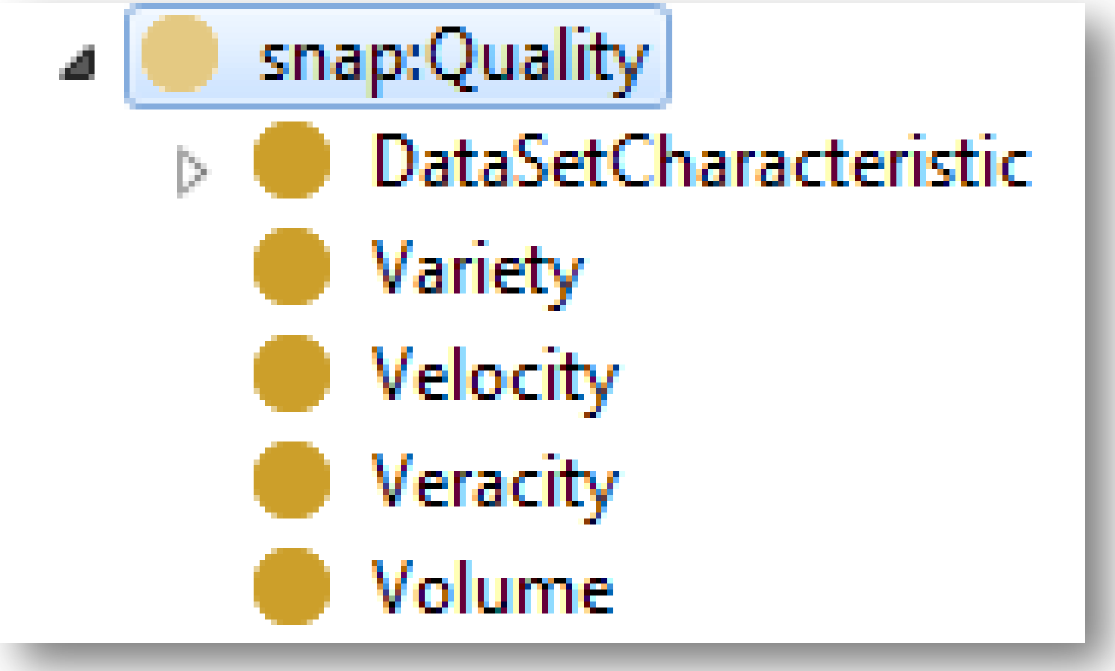
Information Content Entities



- InformationContentEntity
  - AnalyticsResult
  - Characterization
    - WholeHumanGenomeCharacterization
  - Code
  - Comment
  - CorrelationResult
  - Description
  - Label
  - LocationDesignation
    - GridCoordinate
    - LocationAddress
    - Waypoint
  - MeasurementResult
    - CalculationResult
    - FrequencyMeasurementResult
    - HumanGenomeDataMeasurementResult
    - VolumeMeasurementResult
  - Prescription
    - Algorithm
    - Command
    - Directive
    - Order
    - Protocol
    - Standard
  - Question
  - Remark
  - Report
  - Request

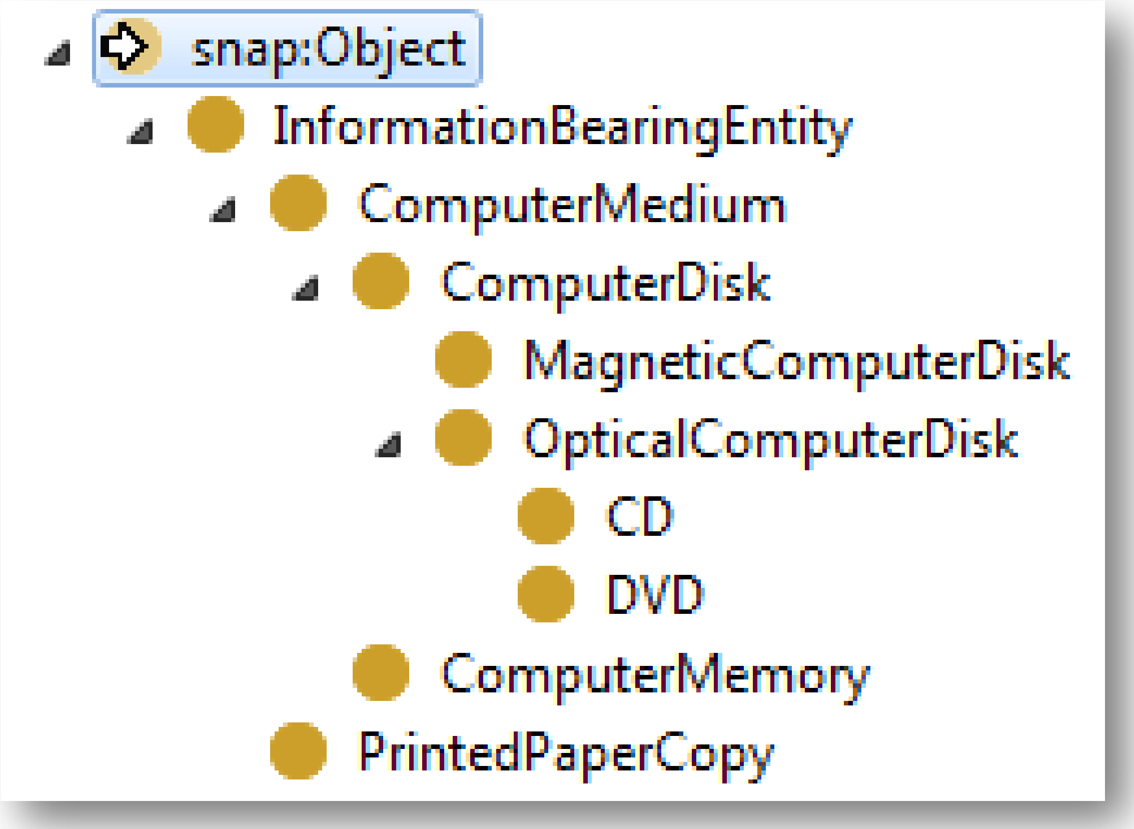
Use Case

Data Characteristics



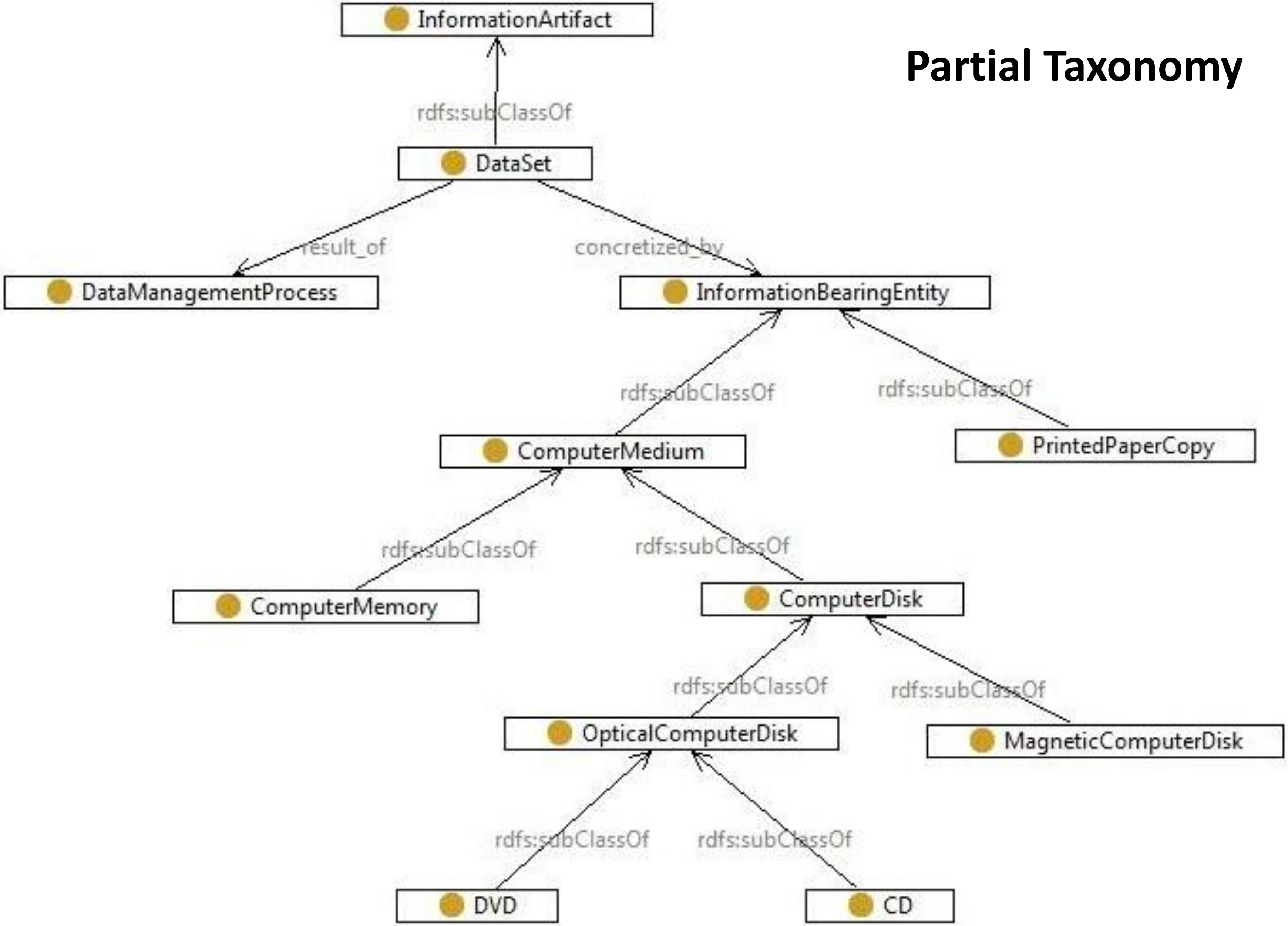
- snap:Quality
  - DataSetCharacteristic
  - Variety
  - Velocity
  - Veracity
  - Volume

Information Bearers

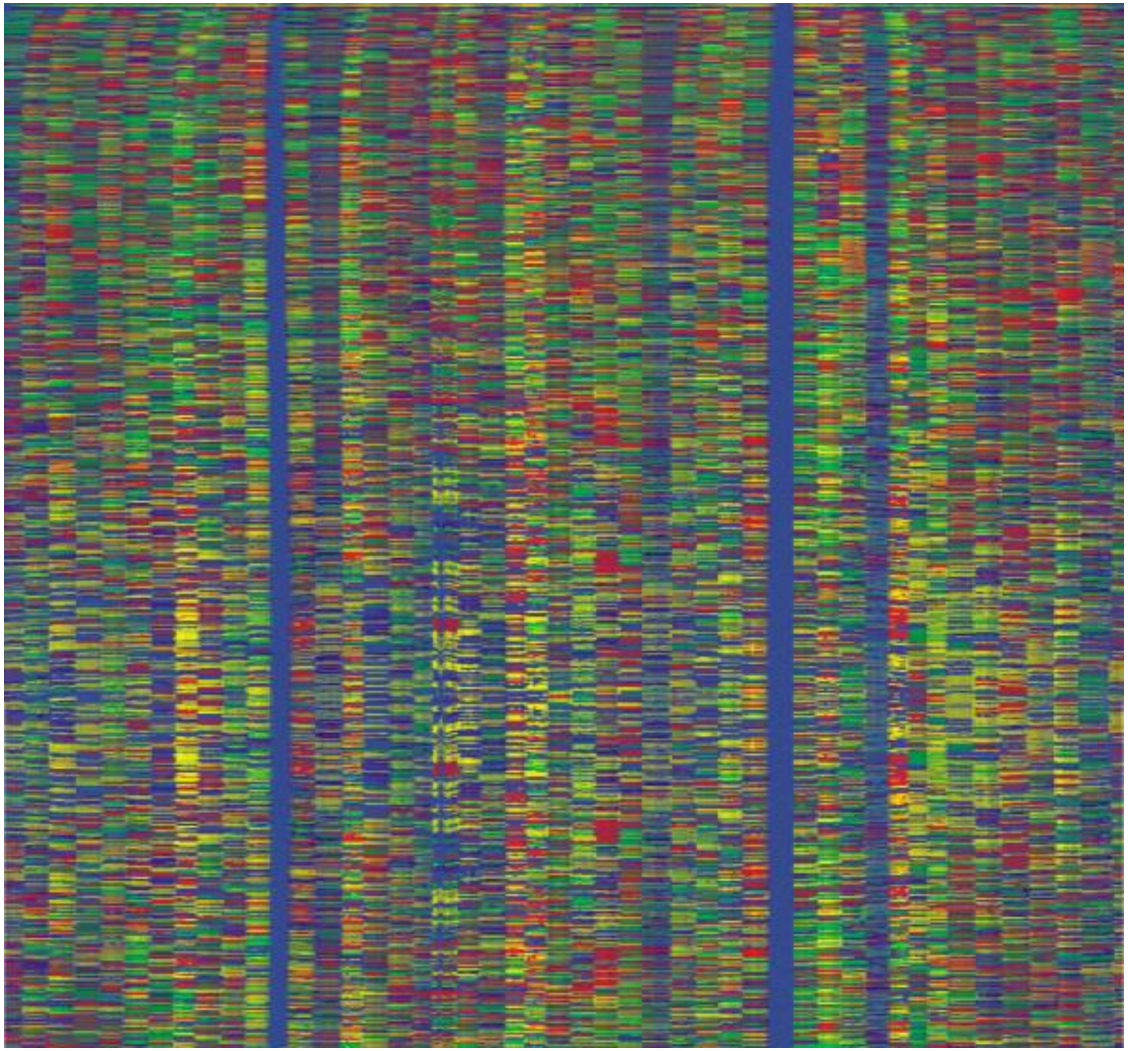


- snap:Object
  - InformationBearingEntity
    - ComputerMedium
      - ComputerDisk
        - MagneticComputerDisk
        - OpticalComputerDisk
          - CD
          - DVD
      - ComputerMemory
    - PrintedPaperCopy

Partial Taxonomy



Human Genome Data



Terms from Human Genome Data Use Case

Use Case Term:	Taxonomical Term:
Genomic Measurements	Genomic Measurement Result (Measurement Result)
Reference Materials	Reference Material Role
Reference Data	Reference Data Role
Reference Methods	Reference Method
Assess Performance	Performance Assessment Process
Genome Sequencing	Genome Sequencing Process
Integrate Data	Data Integration Process
Sequencing Technologies	Data Sequencing Technology (Tool)
Sequencing Methods	Sequencing Method (Process)
Characterization	Characterization (Data Characterization, IA or ICE)
Whole Human Genomes	Whole Human Genome Characterization (IA or ICE?)
Assess Performance	Performance Assessment Process
Genome Sequencing Run	Genome Sequencing Run
Computer System	Computer System
Storage	Data Storage Process
Networking	Computer Networking Process
Processing	Data Processing Process
Software	Software (IAO placement?)
Open Source Sequencing Bioinformatics Software	Bioinformatics Sequencing Software
Data Source	Data Source Role
Sequencer	Sequencer
Volume	Data Volume (Characteristic)
Variety	Data Variety (Characteristic)
Variability	Data Variability (Characteristic)
Veracity	Data Veracity (Characteristic)
Visualization	Data Visualization Process
Data Quality	Data Quality (Characteristic)
Data Types	Data Type
Data Analytics	Data Analytics Process

Terms from Human Genome Data Use Case

Information Artifacts:

Human Genome Data Measurement Result
Characterization (Data Characterization, IA or ICE)
Whole Human Genome Characterization (IA or ICE?)
Performance Assessment
Genome Sequence
Software (IAO placement?)
Data Visualization

Roles and Characteristics:

Reference Material Role
Reference Data Role
Data Source Role
Data Volume (Characteristic)
Data Variety (Characteristic)
Data Variability (Characteristic)
Data Veracity (Characteristic)
Data Visualization Process
Data Quality (Characteristic)

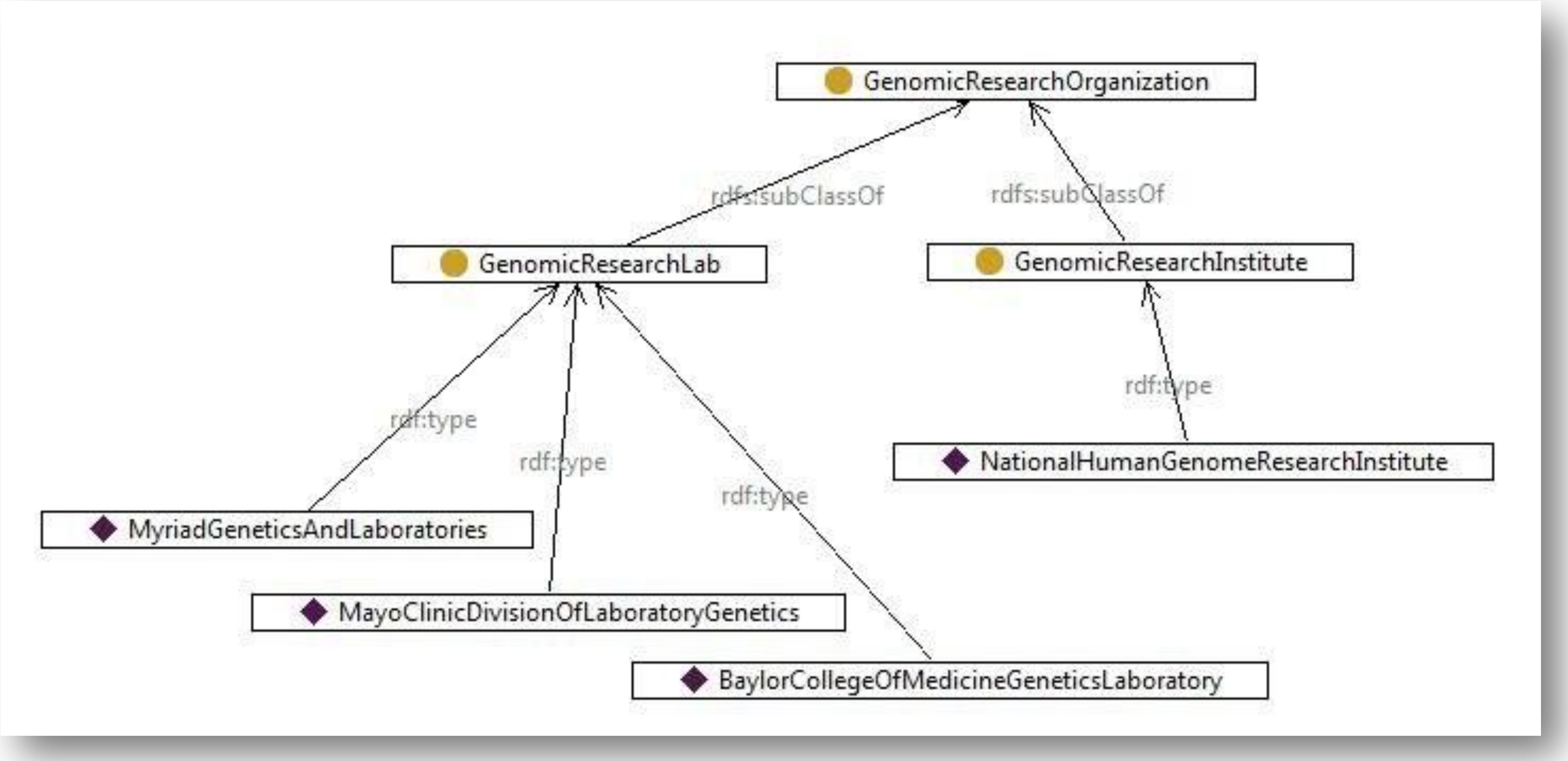
Artifacts/Tools:

Data Sequencing Technology (Tool)
Computer System
Computer Network
Software (IAO placement?)
Bioinformatics Sequencing Software
Sequencer

Processes:

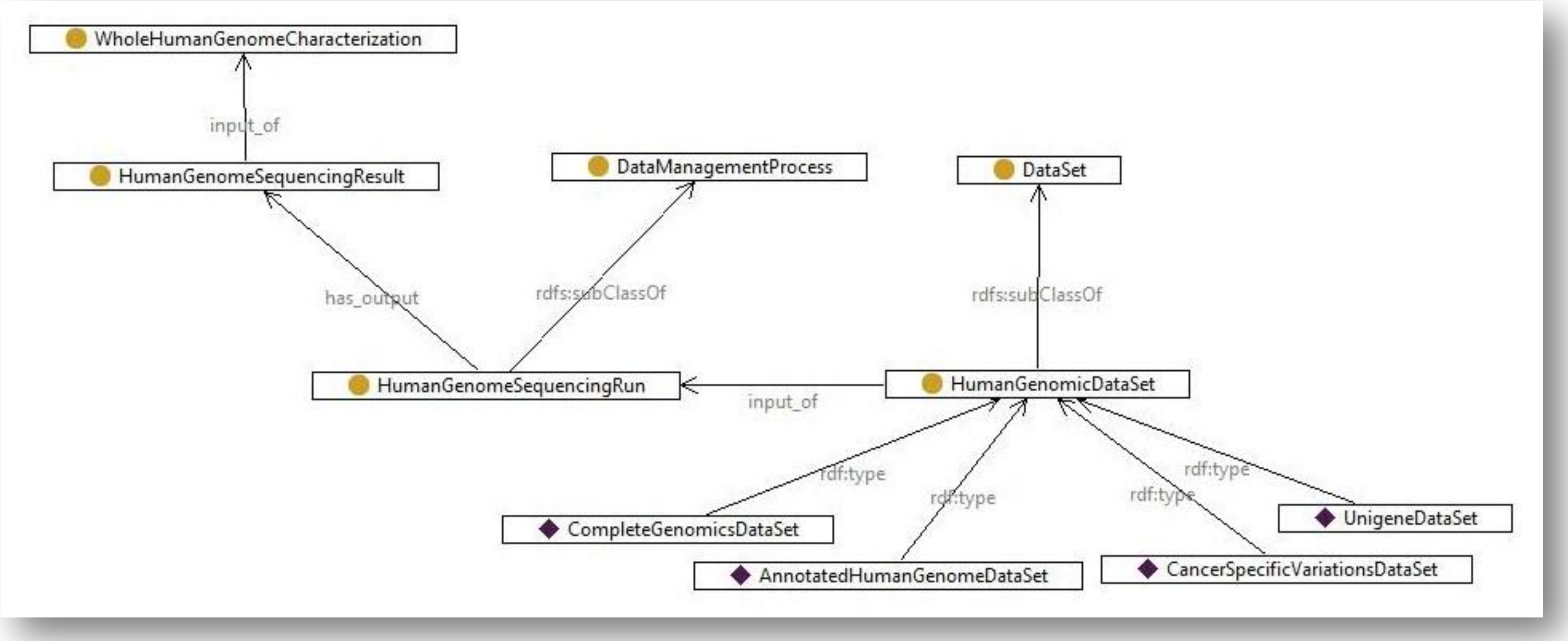
Human Genome Data Measurement Process
Reference Method
Performance Assessment Process
Genome Sequencing Process
Data Integration Process
Sequencing Method (Process)
Data Characterization Process
Performance Assessment Process
Genome Sequencing Run
Data Storage Process
Computer Networking Process
Data Processing Process
Data Visualization Process
Data Analytics Process

Genomic Research Organizations


GenomicResearchOrganization
rdfs:subClassOf
rdfs:subClassOf
GenomicResearchLab
GenomicResearchInstitute
rdf:type
rdf:type
rdf:type
rdf:type
NationalHumanGenomeResearchInstitute
MyriadGeneticsAndLaboratories
MayoClinicDivisionOfLaboratoryGenetics
BaylorCollegeOfMedicineGeneticsLaboratory

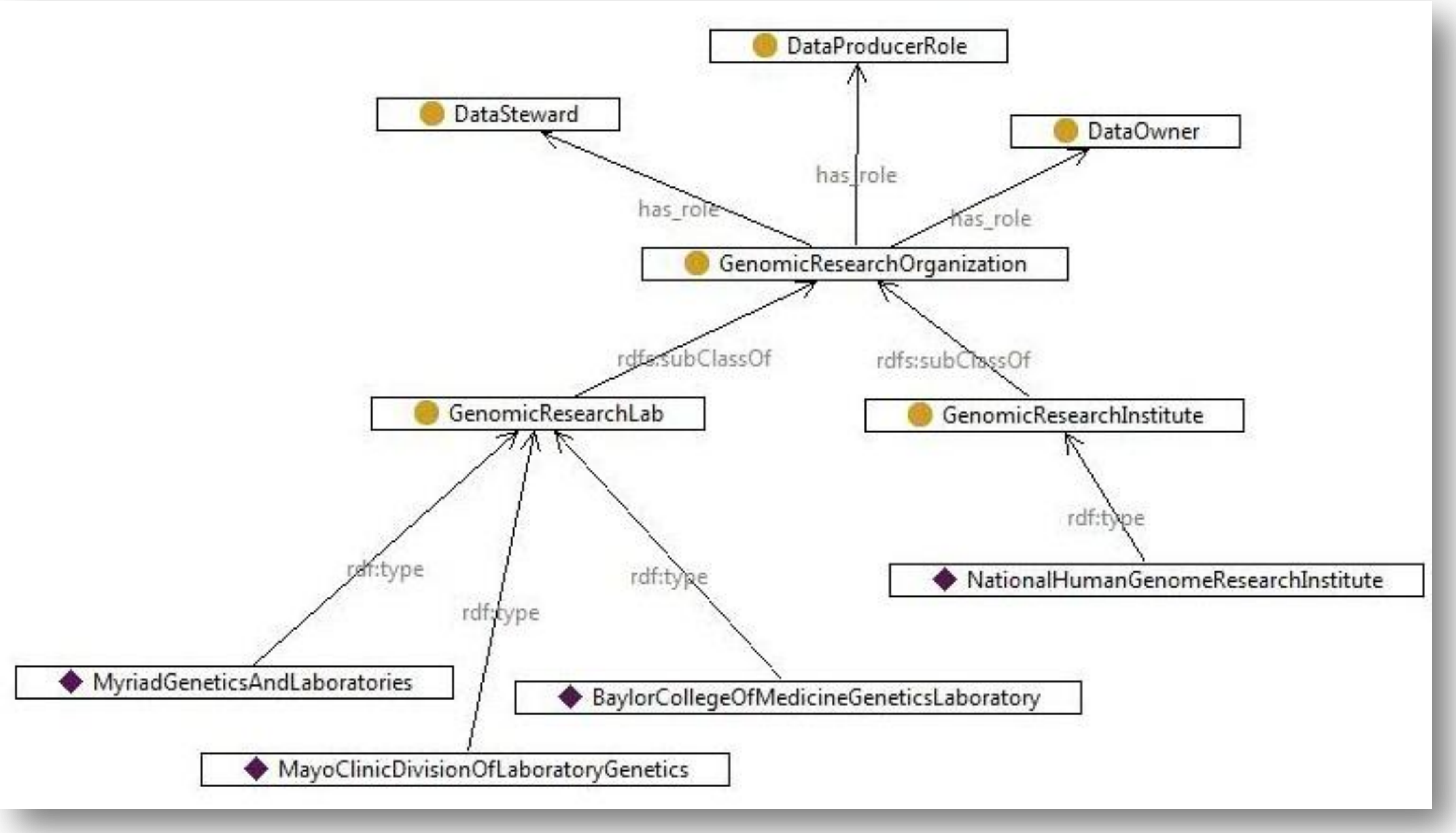
- Instances

DNA Data Sets



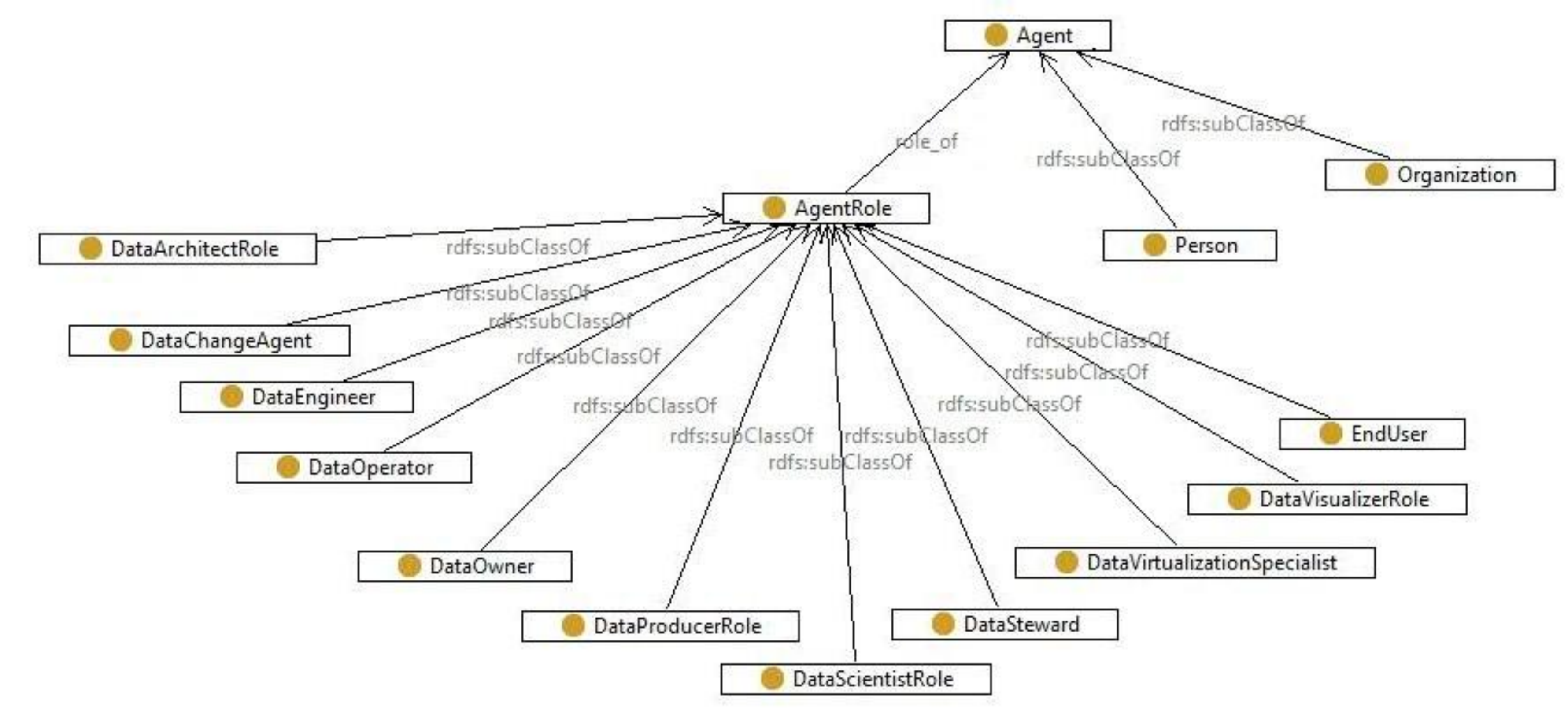
◆ Instances

DNA Organizational Roles

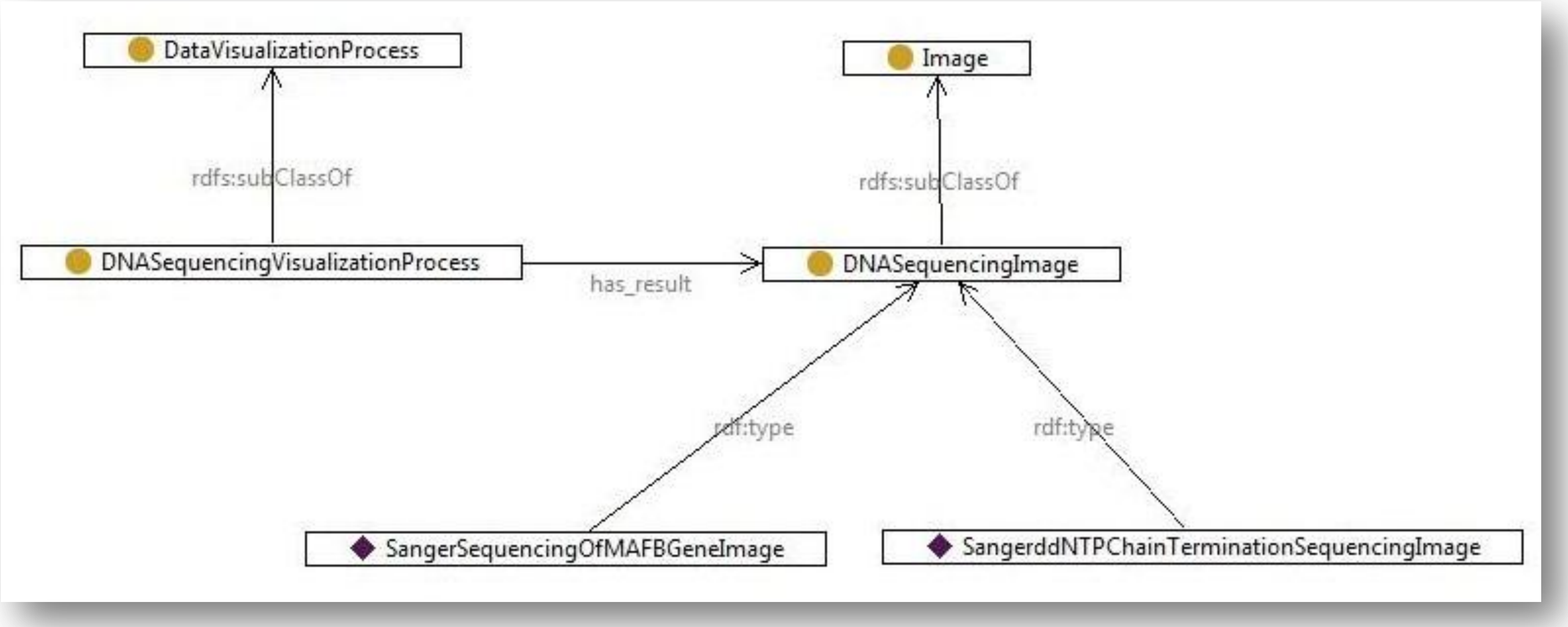

DataProducerRole
DataSteward
DataOwner
has_role
has_role
has_role
GenomicResearchOrganization
rdfs:subClassOf
rdfs:subClassOf
GenomicResearchLab
GenomicResearchInstitute
rdf:type
rdf:type
rdf:type
rdf:type
NationalHumanGenomeResearchInstitute
MyriadGeneticsAndLaboratories
BaylorCollegeOfMedicineGeneticsLaboratory
MayoClinicDivisionOfLaboratoryGenetics

◆ Instances

Agent Roles


Agent
role_of
rdfs:subClassOf
rdfs:subClassOf
Organization
AgentRole
Person
DataArchitectRole
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
DataChangeAgent
DataEngineer
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
EndUser
DataOperator
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
rdfs:subClassOf
DataVisualizerRole
DataOwner
DataVirtualizationSpecialist
DataProducerRole
DataSteward
DataScientistRole

DNA Visualization


DataVisualizationProcess
Image
rdfs:subClassOf
rdfs:subClassOf
DNASequencingVisualizationProcess
has_result
DNASequencingImage
rdf:type
rdf:type
SangerSequencingOfMAFBGeneImage
SangerddNTPChainTerminationSequencingImage

◆ Instances

Conclusion

- This method can be done for any part of the Big Data Taxonomy
- Need SME input for various areas/domains
- Need to add definitions in owl
- Need to expand set of standardized relations
- Link ***instances*** to the taxonomy (e.g. actual data sets, organizations, etc.)