AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

# 4D-2: Data Analytics 2 Data Management

**Alexis B. Carter, MD**
Children's Healthcare of Atlanta

Clinical Informatics Board Review Course

# Clinical Informatics Subspecialty Delineation of Practice (CIS DoP)

**Domain 1: Fundamental Knowledge and Skills (no Tasks are associated with this Domain which is focused on fundamental knowledge and skills)**

**Clinical Informatics**

K001. The discipline of informatics (e.g., definitions, history, careers, professional organizations)
K002. Fundamental informatics concepts, models, and theories
K003. Core clinical informatics literature (e.g., foundational literature, principle journals, critical analysis of literature, use of evidence to inform practice)
K004. Descriptive and inferential statistics
K005. Health Information Technology (HIT) principles and science
K006. Computer programming fundamentals and computational thinking
K007. Basic systems and network architectures
K008. Basic database structure, data retrieval and analytics techniques and tools
K009. Development and use of interoperability/exchange standards (e.g., Fast Health Interoperability Resources [FHIR], Digital Imaging and Communications in Medicine [DICOM])
K010. Development and use of transaction standards (e.g., American National Standards Institute X12)
K011. Development and use of messaging standards (e.g., Health Level Seven [HL7] v2)
K012. Development and use of ancillary data standards (e.g., imaging and Laboratory Information System[LIS])
K013. Development and use of data model standards
K014. Vocabularies, terminologies, and nomenclatures (e.g., Logical Observation Identifiers Names and Codes [LOINC], Systematized Nomenclature of Medicine --Clinical Terms [SNOMED-CT], RxNorm, International Classification Of Diseases[ICD], Current Procedural Terminology [CPT])
K015. Data taxonomies and ontologies
K016. Security, privacy, and confidentiality requirements and practices
K017. Legal and regulatory issues related to clinical data and information sharing
K018. Technical and non-technical approaches and barriers to interoperability
K019. Ethics and professionalism

**The Health System**

K020. Primary domains of health, organizational structures, cultures, and processes (e.g., health care delivery, public health, personal health, population health, education of health professionals, clinical research)
K021. Determinants of individual and population health
K022. Forces shaping health care delivery and considerations regarding health care access
K023. Health economics and financing
K024. Policy and regulatory frameworks related to the healthcare system
K025. The flow of data, information, and knowledge within the health system

**Domain 2: Improving Care Delivery and Outcomes**

K026. Decision science (e.g., Bayes theorem, decision analysis, probability theory, utility and preference assessment, test characteristics)
K027. Clinical decision support standards and processes for development, implementation, evaluation, and maintenance
K028. Five Rights of clinical decision support (i.e., information, person, intervention formats, channel, and point/time in workflow)
K029. Legal, regulatory, and ethical issues regarding clinical decision support
K030. Methods of workflow analysis
K031. Principles of workflow re-engineering
K032. Quality improvement principles and practices (e.g., Six Sigma, Lean, Plan-Do-Study-Act [PDSA] cycle, root cause analysis)
K033. User-centered design principles (e.g., iterative design process)
K034. Usability testing
K035. Definitions of measures (e.g., quality performance, regulatory, pay for performance, public health surveillance)
K036. Measure development and evaluation processes and criteria
K037. Key performance indicators (KPIs)
K038. Claims analytics and benchmarks
K039. Predictive analytic techniques, indications, and limitations
K040. Clinical and financial benchmarking sources (e.g., Gartner, Healthcare Information and Management Systems Society [HIMSS] Analytics, Centers for Medicare and Medicaid Services [CMS], Leapfrog)
K041. Quality standards and measures promulgated by quality organizations (e.g., National Quality Forum [NQF], Centers for Medicare and Medicaid Services [CMS], National Committee for Quality Assurance [NCQA])
K042. Facility accreditation quality and safety standards (e.g., The Joint Commission, Clinical Laboratory Improvement Amendments [CLIA])
K043. Clinical quality standards (e.g., Physician Quality Reporting System [PQRS], Agency for Healthcare Research and Quality [AHRQ], National Surgical Quality Improvement Program [NSQIP], Quality Reporting Document Architecture [QRDA], Health Quality Measure Format [HQMF], Council on Quality and Leadership [CQL], Fast Health Interoperability Resources [FHIR] Clinical Reasoning)
K044. Reporting requirements
K045. Methods to measure and report organizational performance
K046. Adoption metrics (e.g., Electronic Medical Records Adoption Model [EMRAM], Adoption Model for Analytics Maturity [AMAM])
K047. Social determinants of health
K048. Use of patient-generated data
K049. Prediction models
K050. Risk stratification and adjustment
K051. Concepts and tools for care coordination
K052. Care delivery and payment models

**Domain 3: Enterprise Information Systems**

K053. Health information technology landscape (e.g., innovation strategies, emerging technologies)
K054. Institutional governance of clinical information systems
K055. Information system maintenance requirements
K056. Information needs analysis and information system selection
K057. Information system implementation procedures
K058. Information system evaluation techniques and methods
K059. Information system and integration testing techniques and methodologies
K060. Enterprise architecture (databases, storage, application, interface engine)
K061. Methods of communication between various software components
K062. Network communications infrastructure and protocols between information systems (e.g., Transmission Control Protocol/Internet Protocol [TCP/IP], switches, routers)
K063. Types of settings (e.g., labs, ambulatory, radiology, home) where various systems are used
K064. Clinical system functional requirements
K065. Models and theories of human-computer (machine) interaction (HCI)
K066. HCI evaluation, usability engineering and testing, study design and methods
K067. HCI design standards and design principles
K068. Functionalities of clinical information systems (e.g., Electronic Health Records [EHR], Laboratory Information System [LIS], Picture Archiving and Communication System [PACS], Radiology Information System [RIS] vendor-neutral archive, pharmacy, revenue cycle)
K069. Consumer-facing health informatics applications (e.g., patient portals, mobile health apps and devices, disease management, patient education, behavior modification)
K070. User types and roles, institutional policy and access control
K071. Clinical communication channels and best practices for use (e.g., secure messaging, closed loop communication)
K072. Security threat assessment methods and mitigation strategies
K073. Security standards and safeguards
K074. Clinical impact of scheduled and unscheduled system downtimes
K075. Information system failure modes and downtime mitigation strategies (e.g., replicated data centers, log shipping)
K076. Approaches to knowledge repositories and their implementation and maintenance
K077. Data storage options and their implications
K078. Clinical registries
K079. Health information exchanges
K080. Patient matching strategies
K081. Master patient index
**K082. Data reconciliation**
K083. Regulated medical devices (e.g., pumps, telemetry monitors) that may be integrated into information systems
K084. Non-regulated medical devices (e.g., consumer devices)
K085. Telehealth workflows and resources (e.g., software, hardware, staff)

**Domain 4: Data Governance and Data Analytics**

K086. Stewardship of data
K087. Regulations, organizations, and best practice related to data access and sharing agreements, data use, privacy, security, and portability
K088. Metadata and data dictionaries
**K089. Data life cycle**
K090. Transactional and reporting/research databases
K091. Techniques for the storage of disparate data types
K092. Techniques to extract, transform, and load data
**K093. Data associated with workflow processes and clinical context**
**K094. Data management and validation techniques**
K095. Standards related to storage and retrieval from specialized and emerging data sources
K096. Types and uses of specialized and emerging data sources (e.g., imaging, bioinformatics, internet of things (IoT), patient-generated, social determinants)
K097. Issues related to integrating emerging data sources into business and clinical decision making
**K098. Information architecture**
K099. Query tools and techniques
K100. Flat files, relational and non-relational/NoSQL database structures, distributed file systems
K101. Definitions and appropriate use of descriptive, diagnostic, predictive, and prescriptive analytics
K102. Analytic tools and techniques (e.g., Boolean, Bayesian, statistical/mathematical modeling)
K103. Advanced modeling and algorithms
K104. Artificial intelligence
K105. Machine learning (e.g., neural networks, support vector machines, Bayesian network)
K106. Data visualization (e.g., graphical, geospatial, 3D modeling, dashboards, heat maps)
K107. Natural language processing
K108. Precision medicine (customized treatment plans based on patient-specific data)
K109. Knowledge management and archiving science
K110. Methods for knowledge persistence and sharing
K111. Methods and standards for data sharing across systems (e.g., health information exchanges, public health reporting)

**Domain 5: Leadership and Professionalism**

K112. Environmental scanning and assessment methods and techniques
K113. Consensus building, collaboration, and conflict management
K114. Business plan development for informatics projects and activities (e.g., return on investment, business case analysis, pro forma projections)
K115. Basic revenue cycle
K116. Basic managerial/cost accounting principles and concepts
K117. Capital and operating budgeting
K118. Strategy formulation and evaluation
K119. Approaches to establishing Health Information Technology (HIT) mission and objectives
K120. Communication strategies, including one-on-one, presentation to groups, and asynchronous communication
K121. Effective communication programs to support and sustain systems implementation
K122. Writing effectively for various audiences and goals
K123. Negotiation strategies, methods, and techniques
K124. Conflict management strategies, methods, and techniques
K125. Change management principles, models, and methods
K126. Assessment of organizational culture and behavior change theories
K127. Theory and methods for promoting the adoption and effective use of clinical information systems
K128. Motivational strategies, methods, and techniques
K129. Basic principles and practices of project management
K130. Project management tools and techniques
K131. Leadership principles, models, and methods
K132. Intergenerational communication techniques
K133. Coaching, mentoring, championing and cheerleading methods
K134. Adult learning theories, methods, and techniques
K135. Teaching modalities for individuals and groups
K136. Methods to assess the effectiveness of training and competency development
K137. Principles, models, and methods for building and managing effective interdisciplinary teams
K138. Team productivity and effectiveness (e.g., articulating team goals, defining rules of operation, clarifying individual roles, team management, identifying and addressing challenges)
K139. Group management processes (e.g., nominal group, consensus mapping, Delphi method)

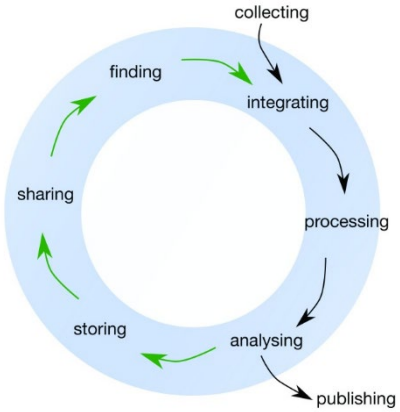
# Knowledge Statements from the DoP

- Data
  - K089. Data life cycle
  - K093. Data associated with workflow processes and clinical context
  - K094. Data management and validation techniques
  - K082. Data reconciliation
  - K098. Information architecture

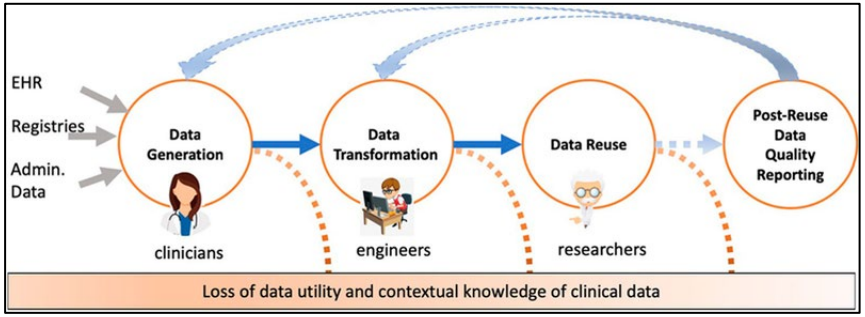
# K089. Data life cycle

# Data Life Cycle

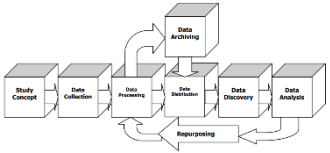
- Definition of this depends on purpose and who you read
- Research
- Content
- Clinical




https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6069748/



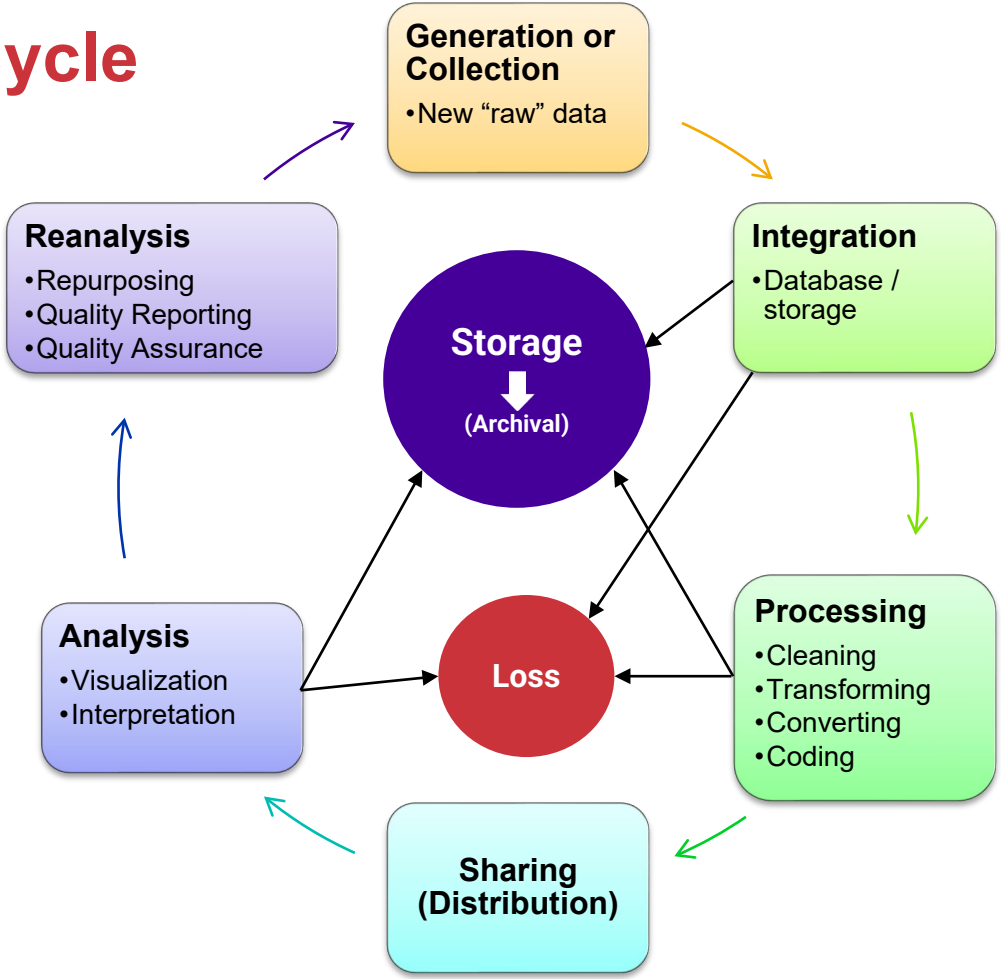

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7133741/



http://opendatafoundation.org/ddi/srg/Papers/DDIModel_v_4.pdf


https://hdsr.mitpress.mit.edu/pub/577rq08d/release/3

# Data Life Cycle



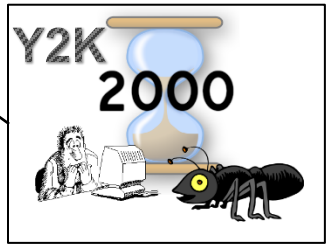
# K094. Data management and validation techniques

# Data Management

- **Data Management**
  - Architectures, practices and procedures for proper management of data lifecycle
- **Data Generation**
  - Generated data needed for care must have a place to be stored in source and target systems
  - Validation of data generation
    - Ensuring that the software or system is performing correctly
    - Both should be documented and stored for later retrieval as needed
    - **Validation**
      - Performing an extensive set of tests (on software) to ensure that it is functioning as expected
      - Test all functions including ones not supposed to be impacted
      - Used on non-FDA approved software and systems
    - **Verification**
      - Performing a limited set of tests (on software) to ensure that it is functioning as expected
      - Tests a sampling of functions including some not supposed to be impacted
      - Used to make sure the FDA-approved software or system did not break in transit or installation

# Data Management

- **Data storage**
  - Storage must be adequate and appropriate for the data being stored
  - Lack of sufficient or appropriate storage can result in errors and system crashes
    - Small scale →
      - Ensuring sufficient element space (MRN length is not > character limits of MRN field)
      - Ensuring that target fields are of the appropriate type (string, numeric, image, etc.)
    - Large scale → ensuring that there are not record limits, date limits or column limits that could be exceeded
  - Validation of data storage integrity
    - Ensures that stored data can be retrieved and used as expected
    - Needs to occur anytime you recover from a downtime
    - Example: have a set of patients where you have known historical data
      - Check for accuracy



# Data Transfer

- **Data transfer**
  - Transferring data from one system (**source system**) to another (**target system**)
    - Methods
      - Manual
      - Flat-file (electronic import or export; may be manually or automatically triggered)
      - Real-time (e.g., HL7, ASTM)
  - **Data transfer validation**
    - Checking the accuracy of data in the source and target systems
    - Check for truncated and missing data
      - Example: Some target systems may not remove or handle HL7 function characters (e.g., tilde ~) that can be embedded in notes and reports. Any data after the tilde will not display to the end-user (truncation).
    - Check for unexpected conversions of data

# Data Transfer

## Data Mapping

- Setting up definitions to map data from source system to target system
  - Enables automated transfer of data between systems
  - Example: mapping medication drug dose in the source system to the correct field in the target system
- Mapping errors
  - Data in the source system is mapped to the wrong field (or to no field) in the target system
  - Mapping quality checks often lacking
  - **Can be invisible to the end-user**

| | |
|---|---|
| Glucose Level | 212 mg/dL |
| Total Cholesterol | 212 mg/dL |

This is actually the patient's cholesterol value

## Types of Mapping Errors

**Human mapping error**

- Can be due to unintentional error, insufficient knowledge or information in subject area to map correctly
- Smaller in scale but can still be very unsafe (e.g., mapping cholesterol values to a glucose field)

**Systematic error**

- Systematic mismapping of data in target system
- Risk is higher when codes are used for mapping
- Risk is higher with conversions of data across an interface
- Can be small (a few items) or very large (all items) in scale

**Errors causing missing data (Data missingness)**

- Completely missing and partially missing (truncated) data
- Mapping to fields with insufficient memory for data (long notes, reports)
- Mapping to non-existent fields

Pfaff E. U4M2L5: Data Validation Importance. AMIA Health Informatics Certification Course. Accessed August 25, 2021.

# Data Transformation

- Converting data into a different format or value through conversion or calculation
- Data transformation is **never** plug and play
  - When performed during data transfer, balance needs for transformation against possible confusion because data in source system will not visually match data in target system
  - Can complicate troubleshooting efforts if not well documented, validated and understood
- Data transformation errors can result in…
  - **Loss of granularity**
    - Source database more detailed than target
    - Discrete data lost or no longer discrete (data rolled up into a single field --> hard to untangle afterward)
  - **Loss of context**
    - Particularly problematic with complex data (e.g., genomics)
    - Example: variant status = pathogenic (can cause disease) BUT…
      - Interpretation = carrier (single heterozygous variant in autosomal recessive disease)

Pfaff E. U4M2L5: Data Validation Importance. AMIA Health Informatics Certification Course. Accessed August 25, 2021.

# Data Validation

- **Data Validation (for interfaced data)**
  - Validating that data from the source system has crossed to the target system…
    - Accurately AND completely AND with the same context/interpretation
  - **Internal validation**: performed by internal healthcare employees
  - **External validation**: performed by external 3rd party services
- When an error is found, always first check source data for accuracy
  - If the source data is wrong, then the transformation will be also
  - If the source data is correct, look for transformation errors

# K082. Data reconciliation

## *(and Data Migration)*

# Data Migration

- Definition
  - Process of transferring data from one storage system or computing environment to another
- Typically refers to large scale transfers (entire database or system)
- When converting from one information system to another, determine what data needs to migrate
  - e.g., LIS conversions require data migration of pathology and blood bank testing history data
- Is a sub-project of a system conversion
  - Can you go live on the new system without the migrated data?
    - i.e., keep the old system up and available for a period of time as view only
  - What date to go live with the migrated data?

# Data Reconciliation

- Definition
  - Technology that uses process information and **mathematical methods** to perform high-level accuracy check of large amounts of migrated data
  - Used with large scale data migrations because detailed validation of all data is not possible
- Purpose
  - Quickly determine whether data migration has gross errors
    - Should be part of the testing and validation process for a data migration before the final migration
    - If data reconciliation passes, then do **data validation** of smaller subsets of data
  - Detects presence or absence of errors, but …
    - Does not usually tell you the source of the error
- Examples
  - Counting total number of records, columns or items between source and target databases
  - Checking $$$ totals of transactions between source and target systems
  - Checking that a column has the same number of non-null values in source and target

# Data Reconciliation Methods

| | |
|---|---|
| **Master Data Reconciliation** | Reconciling only the master data between source and target. Master data is mostly unchanging or slowly changing in nature (e.g., base patient demographics), and no aggregate operation is done on the data set<br>•Total number of items/rows in source vs. target for a particular table<br>•Total number of items/rows based on a specific condition:<br>○ e.g., Number of active users vs. inactive users |
| **Accuracy of Activity** | Ensures that transactions are<br>• Valid<br>• Correct in purpose<br>• Properly authorized |
| **Transactional Data Reconciliation** | Detects mismatches in transactions that can affect business intelligence (BI) reports.<br>•e.g., sum of total revenue calculated from source and target<br>•e.g., Sum of entire item sold, calculated from source and target |
| **Automated Data Reconciliation** | Automated method of reconciling source and target. Data is loaded into meta tables which run data reconciliation algorithms. If the data passes, it is uploaded into the target database. |

# Data Reconciliation and Data Validation

| **Type of data transfer** | **Scale** (amount of data) | **Frequency of transfer** | **Method(s) required** |
|---|---|---|---|
| Continuously interfaced data | Small | Real-time or very frequent | **Data Validation** *(prior to go live and with each change)* |
| **Data migration** | Large | One time or in several large batches | **Data Reconciliation**<br><br>…followed by **data validation of small data subsets** |

https://www.guru99.com/what-is-data-reconciliation.html
https://dwbi.org/pages/12
https://en.wikipedia.org/wiki/Data_validation_and_reconciliation

# Data Migration Errors

- More common when data significantly transformed between source and target
- Small errors → big problems in data interpretation and patient safety
- Can result in
  - Missing or duplicated records
  - Missing, truncated or incorrect values
  - Poor formatting and indecipherability
  - Broken relationships across tables
- Mitigation strategies
  - Allow adequate time and resources for planning, reconciling and validating in batches
  - Migrate only what is necessary (no junk data)
  - Have a backout plan and recovery plan

| Type of Error | Examples |
|---|---|
| **Random errors**<br><br>Errors seen in some but not all of the same set of data | • Run time failures<br>• Network dropouts<br>• Data with unusual contents causing errors (e.g., excessive length, unusual/functional characters) |
| **Systematic errors**<br><br>Errors systematically present in a particular set of data | • Constraint violations (can also result in truncated data) → data fields in target not appropriate for mapped source data (type, size)<br>• Unhandled systematic exceptions<br>• Logical issues or inherent flaws in program |

# K093. Data associated with workflow processes and clinical context

# DIKW Model

- Continuum:
  - **Data** → **Information** → **Knowledge** → **Wisdom**
    - Ideal to make patient care decisions based on Wisdom [Cato et al. 2020]

| | |
|---|---|
| **Wisdom** | Understanding and internalization of knowledge to apply it appropriately |
| **Knowledge** | Derived by discovering patterns and relationships between pieces of information |
| **Information** | Data plus meaning |
| **Data** | Item or signal with little to no meaning by itself |

- Data flows according to communication theory
  - Information source → Sender → *(external factors and noise)* → Receiver → Destination
  - Goal is to minimize external factors and noise
- **Closed system**: Data goes in but does not leave
- **Open system**: Data goes in and can be exported out (data exchange)

*Hardy L. U1M4L2. The Health System – Lecture 2. AMIA Health Informatics Certification Course. Accessed August 25, 2021.*

# Data, workflow processes and clinical context

- See lecture on workflow, workflow re-engineering and change management
- **Workflow** = People + Objects + Information (Data)
- **Context** is critical to how someone uses data in workflow
  - Part of implementation science
  - Roles in innovation adoption, quality improvement, research utilization
  - Squires et al 2019 looked at context attributes which were barriers or enablers to healthcare professionals' use of research evidence in clinical practice – Top 4 (>90% interviews)

| | |
|---|---|
| **1 Resource access** | Having adequate time, access to guidelines / documentation |
| **2 Work structure** | Arrangement of tasks, responsibilities and resources within and between teams |
| **3 Patient characteristics** | Attributes of patients under care |
| **4 Professional role** | Set of expectations associated with a particular clinical occupation |

# K098. Information architecture

# Information architecture

- Covering at a high level → details in Human Computer Interaction lecture
  - Databases and normalization, data warehouses and data marts
- **Information** Architecture
  - a.k.a. information organization and delivery
  - Synonyms
    - Usability engineering, content management, content strategy, user experience (UX) design and interaction design (IxD) https://computer.howstuffworks.com/information-architecture.htm#pt3
  - Definition
    - The way in which content is organized, structured, and labeled for the purposes of helping users find information and complete tasks in an effective and sustainable way
  - https://www.usability.gov/what-and-why/information-architecture.html
  - Different from system/network architecture

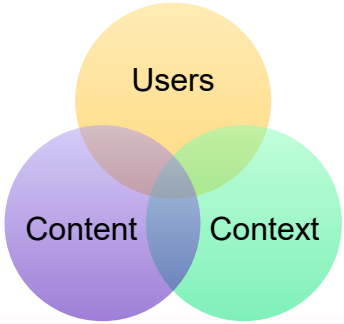
*Pfaff E. U4M2L7: Data and Information Architecture. AMIA Health Informatics Certification Course. Accessed August 20, 2021.*

# Information Architecture

- Raw Data + Context = Information
- Main components

| | |
|---|---|
| **Raw data** | GRCh37-chr7-140453136-T-A |
| **Some context** | *BRAF* p.Val600Glu (V600E) |
| **More context** | *BRAF* p.Val600Glu (V600E) present in pt's melanoma |

| | |
|---|---|
| **Organization schemes and structures** | How information is categorized and structured |
| **Labeling systems** | How data is represented (terminology, ontology) |
| **Navigation systems** | How users browse or move through information |
| **Search systems** | How users look for information |




| | |
|---|---|
| **Context** | Goals, funding, politics, culture, technology, specialty, resources, constraints |
| **Content** | Data and content you want to display |
| **Users** | Users and their frame of reference, needs, assigned tasks, behavior in seeking information, experience |

https://computer.howstuffworks.com/information-architecture.htm#pt3

# That’s a wrap!