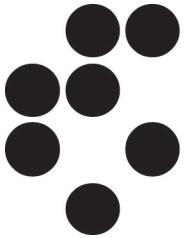


Institut
"Jožef Stefan"
Ljubljana, Slovenija

Knowledge and data representation FAIR data principles

FAIR multi-label clasification tutorial – Part 1

Dragi Kocev, Jasmin Bogatinovski, Ana Kostovska, **Panče Panov**

ECML PKDD 2021

17. 9. 2021

ECML PKDD 2021

Outline

- Basic concepts from knowledge representation
- Data representation and metadata
- FAIR data principles
- TRUST principles for building digital repositories

BASIC CONCEPTS FROM KNOWLEDGE REPRESENTATION

Knowledge-based system and knowledge base

- Representing domain knowledge
  - We need to represent somehow the knowledge about the domain of interest
  - Knowledge must be in a format that can be processed by computers
- In order to operate, a knowledge-based system maintains the knowledge base
- Knowledge base
  - Stores the symbols of the computer model of knowledge in the form of statements about the domain of interest
  - Performs reasoning by manipulating these symbols
- Reasoners are software items that can justify its decisions on issues in a particular domain of interest by querying the knowledge base

Knowledge representation and reasoning

- Knowledge representation and reasoning
  - A branch of artificial intelligence
  - It aims to design computer systems that infer on the basis of a machine-interpretable representation of the world
  - The process of machine reasoning tries to be similar to human reasoning
- Knowledge - based systems
  - They contain a computer model of the domain of interest
  - The symbols serve as a substitute for real-world artifacts
  - Example: physical objects, events, relationships, etc.
- Domain of interest
  - Any part of the real world
  - Any hypothetical system whose knowledge we want to represent

Computer-based knowledge representation

- Knowledge-based systems use computer-based knowledge representation
- Knowledge includes statements of interest
- The represented knowledge can be used to answer questions about the domain of interest
  - With the help of the set of made statements and
  - By using automatic deduction
- A knowledge-based system can draw its own conclusions.

Forms of knowledge representation

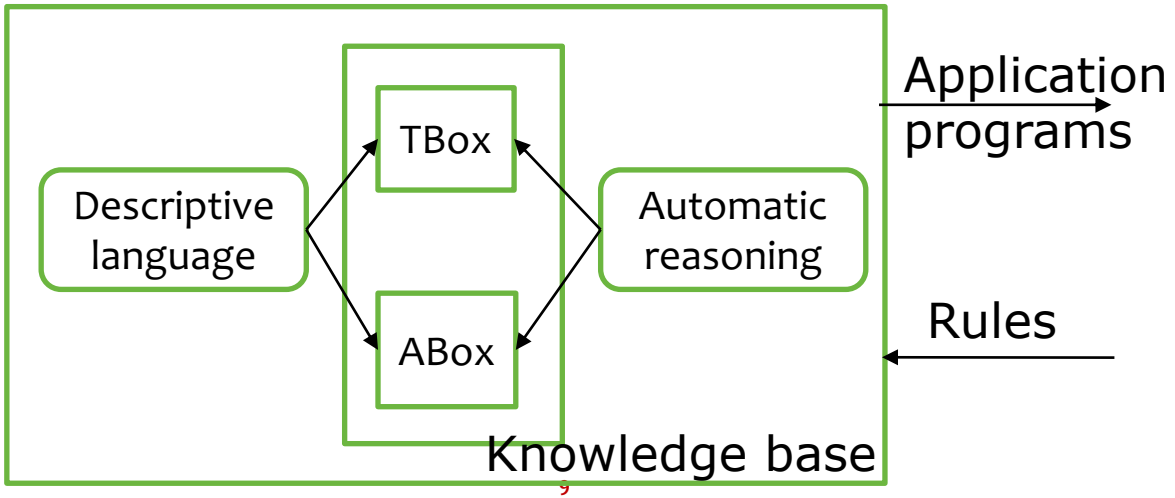
- The representation of knowledge exists in various forms
- Semantic networks
  - They show the taxonomic structure of objects and the relationships between objects
  - Example: Resource Description Framework (RDF) graphs
- Rules
  - The rules reflect the notion of consequence
  - The format of the rules is IF-THEN (IF-THEN), followed by an action to be taken if the conditions are met
- Logic
  - It is used to implement a precise semantic interpretation of the case of semantic networks and rules
  - Logic-based formalisms lay the groundwork for automatic deduction

Reasoning about knowledge

- Computers process the knowledge stored in the knowledge base
  - derive new statements that follow from already saved statements
- The basic operations that can be performed on a knowledge-based system are:
  - tell-operation
    - adds a new statement to the knowledge base
  - ask-operation
    - used to inquire about things that are already in the knowledge base
- Explicit and implicit knowledge
  - the statements added through the "tell" operation contain explicit knowledge of the domain of interest
  - drawing conclusions about the domain of interest by deriving implicit knowledge by processing explicit knowledge
- Inference procedures are implemented as computer programs that perform automatic reasoning
  - computational reasoners

Example:Knowledge-base system based on DL

- Description logics (DL) is a family of logical theories and represent decidable segments of first-order logic.
  - One of the most commonly used paradigms for knowledge representation
  - Descriptive logic contains statements about concepts, individuals and the relationships between them.
- The knowledge base, based on descriptive logic, consists of two components
  - TBox - introduces the terminology of the domain of interest
  - Abox - contains assertions about individuals, expressed with the help of a dictionary from TBox



Ontologies as representational artefacts in computer science

- Ontologies in computer science
  - Conceptual models of what 'exists' in the domain of interest
  - Use the format that can be interpreted by computers
  - Models use knowledge representation techniques
- Using the ideas of ontological categories from philosophy:
  - we limit the ontology to cover a small subset of the world (our domain of interest)
- Ontology terms are used to represent the knowledge by making statements about a domain of interest
- The use of ontologies along with knowledge representation formalisms
  - In semantic networks, an ontology provides us with the meaning (semantics) of the nodes and arcs
  - In rules and logical formulas, an ontology provides us with the meaning of predicates and constants

Why are ontologies useful?

- To share a common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

Ontology vocabulary

- The main components (concepts, relations, individuals) of the ontology constitute the ontological vocabulary of the domain of interest
- Ontology can be considered as a set of statements expressed in a ontology vocabulary
- Statements are also called axioms
- Ontologies come in a variety of forms
  - The knowledge engineer looks at the ontology from a graphic perspective
  - For the purpose of storage and reasoning, ontologies are written using ontological languages in a specific format that can be processed by computers
- There are different ontological languages based on different formalisms for the presentation of knowledge

Formal ontologies and Semantic web technologies

- Semantic web technologies and formal ontologies
  - Currently a popular solution to data and knowledge sharing
- Set of technologies to support the knowledge-sharing process (RDF, RDFS, OWL)
- Querying:
  - enables us to retrieve explicit facts from the knowledge base datasets
  - SPARQL query language
- Use of reasoners: inference of additional statements from those given explicitly

Dean Allemang and James Hendler. 2011. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL (2 ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ontology languages

- Resource Description Framework (RDF) in RDF Schema (RDFS)
  - an initial language standardization initiative for the semantic labeling of online web resources
  - introduced by the World Wide Web Consortium W3C
  - are used to store metadata
- Based on RDF (S), the W3C designed the Ontology Web Language (OWL) family of languages
  - They are used to describe ontologies on the semantic web
  - Languages have different expressive power

Knowledge graphs

- Knowledge graph uses a graph-structured data model or topology to integrate data
- Knowledge graphs may make use of ontologies as a schema layer
- They allow logical inference for retrieving implicit knowledge rather than only allowing queries requesting explicit knowledge

Relevant ontological representations for machine learning and data mining

- Variety of ontologies, vocabularies and schemes
- OntoDM-core, OntoDT, OntoDM-KDD (Panov et al, 2014;2016): a set of modular ontologies in the context of a general framework for data mining (Džeroski, 2006)
- Exposè (Vanschoren and Soldatova, 2010) and MEX vocabulary (Esteves et al., 2015): representations of machine learning experiments
- DMOP (Keet et al., 2015): meta-mining of data mining workflows
- ML-schema: an initiative to harmonize the developed ontologies

Panov, P., Soldatova, L., & Džeroski, S. (2014). Ontology of core data mining entities. Data Mining and Knowledge Discovery, 28(5), 1222-1265.

Panov, P., Soldatova, L. N., & Džeroski, S. (2016). Generic ontology of datatypes. Information Sciences, 329, 900-920.

Džeroski, S. (2006) Towards a general framework for data mining. In International Workshop on Knowledge Discovery in Inductive Databases (pp. 259-300). Springer, Berlin, Heidelberg.

Vanschoren, J., & Soldatova, L. (2010). Exposé: An ontology for data mining experiments. In *International workshop on third generation data mining: Towards service-oriented knowledge discovery (SoKD-2010)* (pp. 31-46).

Esteves, D., Moussallem, D., Neto, C. B., Soru, T., Usbeck, R., Ackermann, M., & Lehmann, J. (2015, September). MEX vocabulary: a lightweight interchange format for machine learning experiments. In Proceedings of the 11th International Conference on Semantic Systems (pp. 169-176). ACM.

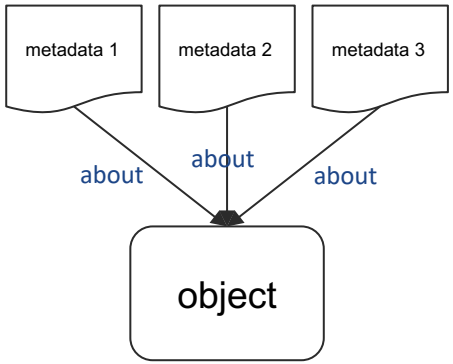
Keet, C. M., Ławrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., ... & Hilario, M. (2015). The data mining optimization ontology. *Journal of web semantics*, *32*, 43-53.

Machine Learning Schema Community Group. URL: https://www.w3.org/community/ml-schema/

DATA REPRESENTATION AND METADATA

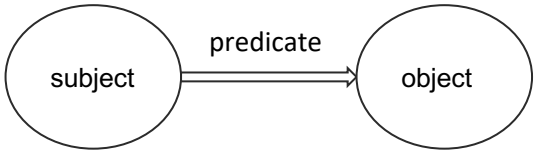
The world of metadata

- Metadata is all around us, all the time.
- When metadata is doing its job well, it just fades in the background, unnoticed and nearly invisible
- Why store data about an object, when you have the object itself?
  - Without the data about the objects contained in a space, any complex space is indistinguishable from chaos.
  - If you want to find the object in timely fashion, you need metadata about it.
- An incredible amount of information about the objects can be inferred from only metadata



What are metadata?

- The word “metadata” indicates something that is beyond the data
  - a statement or statements about the data
- Definition: “Metadata is a statement about a potentially informative object” (Jeffrey Pomerantz)
  - Potentially informative object about which we are making a statement is called a resource
  - The description is what we are saying about the resource
- A statement has 3 parts
  - the subject of our description
  - predicate: a category of relationship between the resource and some other object
  - other object that has the predicated relationship with the resource
- Most adequate representational model to use: RDF
  - A data model according to which most metadata is structured



Why should you use metadata?

- It can make your data discoverable
  - People can perform search based on the metadata
- It can make your data more reusable
  - Because it makes them understandable
  - Facilitates finding related data
- It makes your data more reproducible
  - If you know how/why/where it was collected, it helps others to reproduce your research/experiment in order to validate it

Types of metadata

- Descriptive metadata
  - provides a description of an object
- Administrative metadata
  - provides information about the origin and maintenance of an object
- Structural metadata
  - provides information about how an object is organized
- Preservation metadata
  - provides information necessary to support the process of preserving an object
- Use metadata
  - provides information about how an object has been used

Who should create metadata?

- Ideally the same person or people who created the data
  - they understand the data the best
- People responsible for the data's distribution and curation are also well-placed to add additional metadata
  - This especially holds for structural metadata

What makes good metadata?

- Metadata is good if it allows your data to be found and understood by all those who might want to make use of it
- Complete
- Accurate
- Precise
- Conforming to standards
  - Semantic: Meaning of Terms
  - Which metadata are mandatory
  - Formatting / Syntax
- Accessible
  - Online, addressable (can be linked to), harvestable

Metadata schema

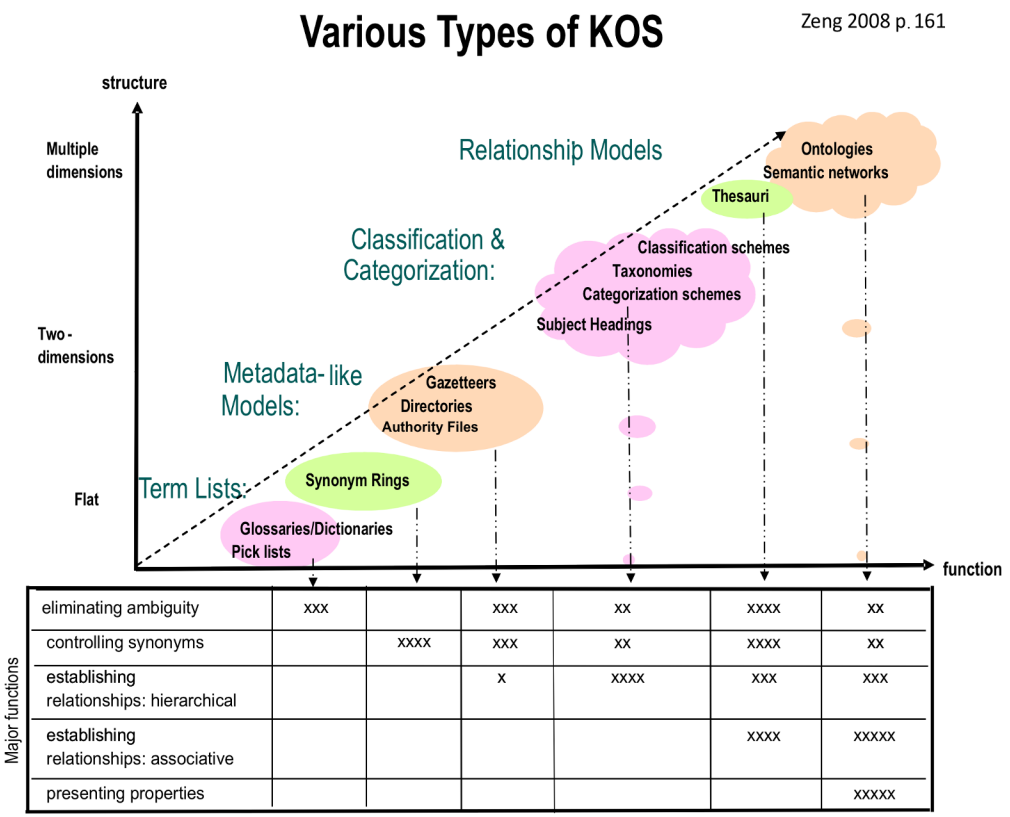
- Metadata schema is a set of rules about what sorts of subject-predicate-object statements (triples) one is allowed to make and how
  - e.g. Dublin Core: metadata schema designed to enable description of any resource
- Predicates are called elements: a category of statement that can be made about a resource
  - Element name denotes an attribute of a resource
  - Value is the data that is assigned to an element
  - Element-value pair: the totality of a single statement about a resource

Important metadata standards

- There are many standards available to document data
- Each has a different focus
- Your choice will depend on:
  - your domain of interest
  - your motivation for using metadata
- Dublin Core Metadata Initiative (https://dublincore.org/)
  - DCMI Metadata Terms
  - Dublin Core Metadata Element Set
- Metadata Encoding and Transmission Standard (METS)
  - http://www.loc.gov/standards/mets/

Knowledge organization systems

- The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management
- Different types of knowledge organization systems
  - Different function
  - Different structure
  - Various level of expressivity and complexity
- Ontologies are most expressive



Source: shorturl.at/kFGOZ

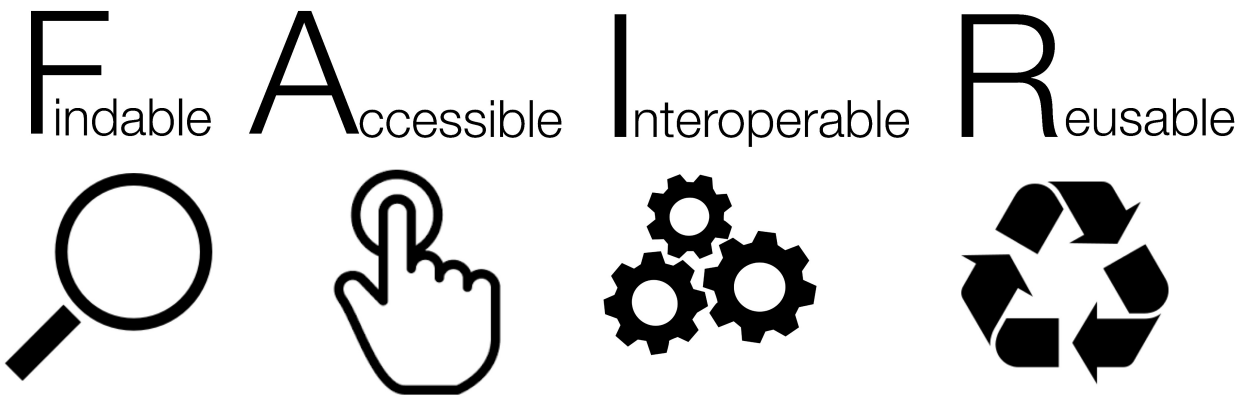
Semantic annotation

- Semantic annotation is the process of tagging documents or any data with relevant concepts originating from ontologies and vocabularies
  - The documents or data are enriched with metadata: references that link the content to concepts, described in an ontology or a knowledge graph
  - This makes unstructured content easier to find, interpret and reuse
- With semantic annotation we want to create references of the data to accepted ontologies
- Semantic annotation can be performed manually or (semi)automatically
- Many of the semantic technologies that can be used to annotate information
  - RDF, RDFS, OWL
  - Querying can be done with SPARQL

FAIR DATA PRINCIPLES

FAIR data principles

- FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability
- FAIR principles focus on enhancing the ability of computers to automatically find, access and (re)use data with minimal human intervention



Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Findability

Metadata and data should be discoverable by both humans and machines

- (Meta)data are assigned a globally unique and persistent identifier (F1)
- Data are described with rich metadata (F2)
- Metadata clearly and explicitly include the identifier of the data it describes (F3)
- (Meta)data are registered or indexed in a searchable resource (F4)

Accessibility

The first step in (re)using data is to find them,
the second is to know how and under which condition they can be accessed

- (Meta)data are retrievable by their identifier using a standardized communications protocol (A1)
  - The protocol is open, free, and universally implementable (A1.1)
  - The protocol allows for an authentication and authorization procedure, where necessary (A1.2)
- Metadata are accessible, even when the data are no longer available (A2)

Interoperabiliy

Data usually need to be integrated with other data, and be interoperable with different applications

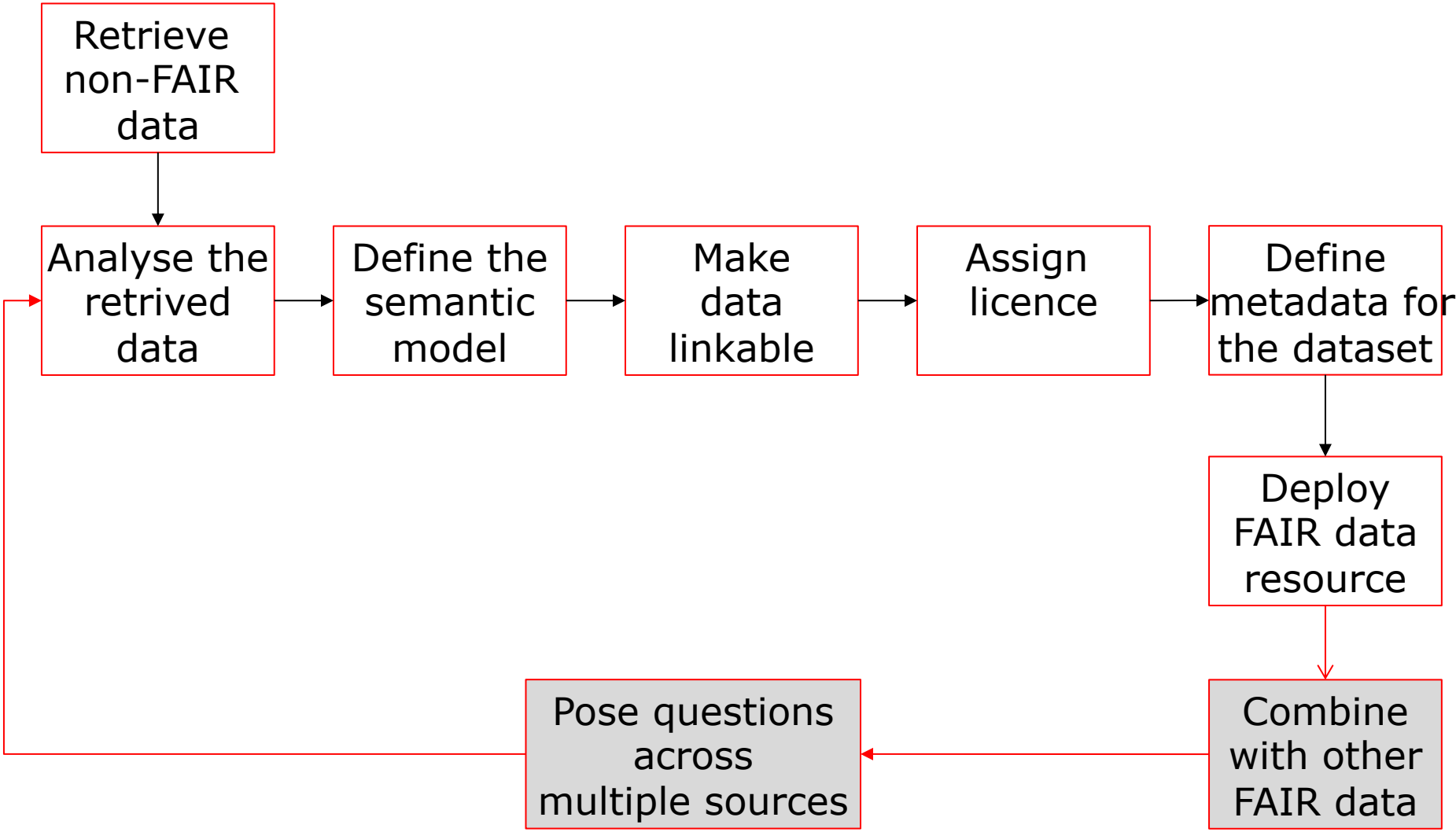
- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (I1)
- (Meta)data use vocabularies that follow FAIR principles (I2)
- (Meta)data include qualified references to other (meta)data (I3)

Reusability

Optimizing the reuse of data is the ultimate goal

- Meta(data) are richly described with a plurality of accurate and relevant attributes (R1)
  - (Meta)data are released with a clear and accessible data usage license (R1.1)
  - (Meta)data are associated with detailed provenance (R1.2)
  - (Meta)data meet domain-relevant community standards (R1.3)

FAIRification Process



https://www.go-fair.org/fair-principles/fairification-process/

Transforming FAIR into practice

- FAIR Cookbook
  - https://fairplus.github.io/the-fair-cookbook/content/home.html
  - Created by researchers and data management professionals
  - Recepies to make your data FAIR
- FAIR evaluation services
  - Perform an automatic assessment of a dataset against the FAIR principles expressed as nanopublications https://w3id.org/AmIFAIR
- Fairsharing.org
  - https://fairsharing.org/
  - A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies

TRUST PRINCIPLES FOR DIGITAL REPOSITORIES

TRUST principles for digital repositories

- "Repositories must earn the trust of the communities they intend to serve and demonstrate that they are reliable and capable of appropriately managing the data they hold." (Lin et al., 2020)
- To make data FAIR whilst preserving them over time requires trustworthy digital repositories (TDRs)
  - Sustainable governance and organizational frameworks
  - reliable infrastructure
  - comprehensive policies supporting community-agreed practices
- Set of guiding principles to demonstrate digital repository trustworthiness
  - TRUST: Transparency, Responsibility, User focus, Sustainability and Technology

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

Transparency

- Mission statement and scope of the repository is clearly stated
- Terms of use are known in advance both for the repository and the data collections
- There exists a minimum digital preservation timeframe for the data collections
- Any additional features or services of the repository are clearly stated

Responsibility

- Adherence to community metadata and curation standards, as well as providing stewardship of the data collection (technical validation, documentation, quality control, authenticity protection, long term persistence)
- Providing data services (user interface, data download, backend server processing)
- Management of intellectual property rights of data producers, protecting of sensitive information, management of security of the system and its content

User focus

- The repository needs to focus on serving its target community
- Enabling community users to find, explore and understand their data collections with respect to potential reuse
- Encouragement for fully described data at submission time and facilitating feedback from users of the data at a later stage
- Enforce the used of metadata schemas, file formats, controlled vocabularies, ontologies and semantics from the user community

Sustainability

- Ensuring uninterrupted access to the data collections for current and future users and communities
- Ability of the repository to provide services over time and to adapt to the user/community needs: adding new services or improving old ones
- Planning and providing long-term preservation of data so that they are discoverable, accessible and usable in the future

Technology

- Repository depends on the interaction of people, processes and technology (software, hardware, technical services)
- Enables the delivery of the TRUST principles
- Implementation of data management and curation standards, tools and technologies
- Planning, detecting and handling of different security events

Summary

- Basic concepts from knowledge representation
  - knowledge-based system, knowledge base, representational model, knowledge modelling, ontologies, semantic web technologies
- Data representation and metadata
  - what is metadata?, describing objects of interest and semantic annotation, types of metadata, metadata schema, metadata standards
- FAIR data principles
  - Findable data, Accessible data, Interoperable data, Reusable data
- TRUST principles for building digital repositories
  - Transparency, Responsibility, User focus, Sustainability, Technology