WELCOME: DAY 2

Session 2

FAIR explained and actionized

FAIR DATA STEWARDSHIP COURSE

TOWARDS A GO FAIR READINESS PROGRAM

FAIR PRINCIPLES: MANY WAYS TO LOOK AT THEM

FAIR PRINCIPLES

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

> A1.1 the protocol is open, free, and universally implementable;
>
> A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

> R1.1. (meta)data are released with a clear and accessible data usage license;
>
> R1.2. (meta)data are associated with detailed provenance;
>
> R1.3. (meta)data meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618

FAIR DATA PRINCIPLES - METADATA

Findable:

F1. metadata are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. metadata are registered or indexed in a searchable resource;

Accessible:

A1. metadata are retrievable by their identifier using a standardized communications protocol;

- A1.1 the protocol is open, free, and universally implementable;
- A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. metadata use vocabularies that follow FAIR principles;

I3. metadata include qualified references to other metadata;

Reusable:

R1. metadata are richly described with a plurality of accurate and relevant attributes;

- R1.1. metadata are released with a clear and accessible data usage license;
- R1.2. metadata are associated with detailed provenance;
- R1.3. metadata meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618

FAIR DATA PRINCIPLES – DATA/DIGITAL RESOURCES

Findable:

F1. data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. metadata are registered or indexed in a searchable resource;

Accessible:

A1. metadata are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. metadata use vocabularies that follow FAIR principles;

I3. metadata include qualified references to other (meta)data;

Reusable:

R1. metadata are richly described with a plurality of accurate and relevant attributes;

R1.1. metadata are released with a clear and accessible data usage license;

R1.2. metadata are associated with detailed provenance;

R1.3. metadata meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618

FAIR DATA PRINCIPLES – SUPPORTING ELEMENTS

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1. the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation;

I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

R1.1. (meta)data are released with a clear and accessible data usage license;

R1.2. (meta)data are associated with detailed provenance;

R1.3. (meta)data meet domain-relevant community standards;

https://www.nature.com/articles/sdata201618

REPOSITORIES SUPPORTING USERS TO ACHIEVE FAIR

Findable:

✓ F1. (meta)data are assigned a globally unique and persistent identifier;

✓ F2. data are described with rich metadata;

✓ F3. metadata clearly and explicitly include the identifier of the data it describes;

✓ F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

✓ A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

✓ A2. metadata are accessible, even when the data are no longer available;

Interoperable:

✓ I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

✓ I2. (meta)data use vocabularies that follow FAIR principles;

✓ I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

✓ R1.1. (meta)data are released with a clear and accessible data usage license;

✓ R1.2. (meta)data are associated with detailed provenance;

✓ R1.3. (meta)data meet domain-relevant community standards;

FAIR PRINCIPLES DETAILED

F1. (meta)data are assigned a globally unique and persistent identifier

- **What does it mean?**
  - We need an identification mechanism, e.g., PID, PURL, DOI, …
  - This mechanism needs to guarantee global uniqueness of the issued identifier, i.e., every time a given identifier is called, the same resource is points to is retrieved
  - This mechanism needs to guarantee persistency of the issued identifier, i.e., what happens when the identifier scheme is changed?
- **What do we need to fulfill this principle?**
  - How to describe the used identification mechanism?
  - How to properly identify the identifier service? I.e., what is the commonly agreed vocabulary that can represent that a given piece of information is the identifier of a digital resource?
  - What is the uniqueness policy?
    - How to represent the policy in a computer-actionable way?
    - What is the required content of the policy, e.g., uniqueness mechanism?
  - What is the persistency policy?
    - How to represent the persistency policy in a computer-actionable way?
    - What is the required content of the policy, e.g., persistency over updates of the mechanism?
  - What is resolved by sending a request to the identifier, the actual digital resource, its metadata, etc.? I.e, what is the protocol for getting the actual digital resource from its identifier?

F2. data are described with rich metadata

- **What does it mean?**
  - If we don’t have the identifier, the digital resource should be described with rich enough metadata that we can find it through the combination of the items in this metadata
- **What do we need to fulfill this principle?**
  - For different types of digital resources, what are the minimal metadata elements that provide this richness?
  - How to describe the metadata in a commonly agreed and computer-actionable way? By using a common way to represent metadata, tools can be made that are able to interpret metadata from any kind of digital resource.

F3. metadata clearly and explicitly include the identifier of the data it describes

- **What does it mean?**
  - The discovery of a digital resource should be possible from its metadata. For this to happen, the metadata must explicitly contain the identifier for the digital resource it describes.
- **What do we need to fulfill this principle?**
  - How to differentiate the information about the digital resource’s identifier and the one about its metadata identifier? I.e., a metadata record contains two identifiers, of itself (the metadata record) and of the data they describe. What is the vocabulary that contains concepts to describe a metadata identifier and digital resource they describe

FAIR PRINCIPLES

F4. (meta)data are registered or indexed in a searchable resource

- **What does it mean?**
  - Most people use a search engine to initiate a search for a particular digital resource. If the resource or its metadata are not index in a searchable resource, the capability for an individual to find it is substantially reduced.
- **What do we need to fulfill this principle?**
  - For the data part, a full indexing is equivalent to allowing complete and direct querying on the data, which may not be feasible every time. An intermediate step would be to select a number of relevant parts of the data to be highlighted by their metadata, which would be indexed. E.g., in a dataset describing gene information, it may be relevant to allow the indexing of the unique genes that the dataset has information about.
  - Search engines benefit from common interfaces (or at least interfaces that are described in a commonly agreed way) to allow the harvesting of the elements (metadata and/or data) to be indexed.
  - A common representation format for the metadata also improves the possibility of different searchable resources to index the metadata records.

FAIR PRINCIPLES

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

- **What does it mean?**
  - In order to access a digital resource, the requestor needs to be able to implement the used communication protocol. Therefore, this protocol should be open, free and universally implementable. Moreover, the protocol should also describe whether authentication and authorization mechanisms are required.
- **What do we need to fulfill this principle?**
  - How to describe the communication (accessibility) protocol?
  - What are the elements required in the description of the communication (accessibility) protocol, including the authentication and authorization procedure?
  - How to demonstrate that the protocol is open, free and universally implementable?

A2. metadata are accessible, even when the data are no longer available

- **What does it mean?**
  - Cross-reference to data from third-party's FAIR data and metadata will naturally degrade over time. Therefore, it is important for FAIR providers to continue to provide descriptors of what the data was to assist in the continued interpretation of those third-party data.
- **What do we need to fulfill this principle?**
  - How to guarantee long-term persistency of the metadata?
  - How to describe that the data (digital resource) referred by the metadata are no longer accessible? Is it necessary to inform why?
  - How to harmonize the persistency of the metadata with the GDPR's "right to be forgotten"?

A2. metadata are accessible, even when the data are no longer available

- **What does it mean?**
  - Cross-reference to data from third-party's FAIR data and metadata will naturally degrade over time. Therefore, it is important for FAIR providers to continue to provide descriptors of what the data was to assist in the continued interpretation of those third-party data.
- **What do we need to fulfill this principle?**
  - How to guarantee long-term persistency of the metadata?
  - How to describe that the data (digital resource) referred by the metadata are no longer accessible? Is it necessary to inform why?
  - How to harmonize the persistency of the metadata with the GDPR's "right to be forgotten"?

FAIR PRINCIPLES

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- **What does it mean?**
  - The digital resource is described using a formal, accessible, shared and broadly applicable language
  -
- **What do we need to fulfill this principle?**
  - How to inform the language used to represent the digital object?
  - How to provide this information for the metadata? In a meta-metadata?
  - How to demonstrate the formality (BNF), accessibility (resolution of the language description document), shareability and broad applicability of the language (IANA media-type?)?

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- **What does it meanC?**
  - The digital resource is described using a formal, accessible, shared and broadly applicable language
  -
- **What do we need to fulfill this principle?**
  - How to inform the language used to represent the digital object?
  - How to provide this information for the metadata? In a meta-metadata?
  - How to demonstrate the formality (BNF), accessibility (resolution of the language description document), shareability and broad applicability of the language (IANA media-type?)?

I2. (meta)data use vocabularies that follow FAIR principles

- **What does it meean?**
  - The metadata values and qualified relations should themselves be FAIR

- **What do we need to fulfill this principle?**
  - Inform which vocabularies are used
  - What is the minimal FAIRness for these vocabularies to be considered to follow FAIR principles?

FAIR PRINCIPLES

I3. (meta)data include qualified references to other (meta)data

- **What does it mean?**
  - Relationships within digital resources and between local and third-party data, have explicit and “useful” semantic meaning
- **What do we need to fulfill this principle?**
  - Qualify (provide proper semantics) the references to other digital resources
  - As per I2, these references (and their qualifiers) should also be FAIR

FAIR PRINCIPLES

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

> **R1.1. (meta)data are released with a clear and accessible data usage license;**
>
> **R1.2. (meta)data are associated with detailed provenance;**
>
> **R1.3. (meta)data meet domain-relevant community standards;**

- **What does it mean?**
  - Digital resources should inform who has which rights under which circumstances (license), what is their provenance and use relevant standards adopt by the community in which the resource has been created/used
- **What do we need to fulfill this principle?**
  - Inform the usage license:
    - What representation format can be used for a computer-actionable license description?
    - What are the required concerns that should be present in this description (rights, conditions, …)?

FAIR PRINCIPLES

R1. (meta)data are richly described with a plurality of accurate and relevant attributes;

R1.1. (meta)data are released with a clear and accessible data usage license;

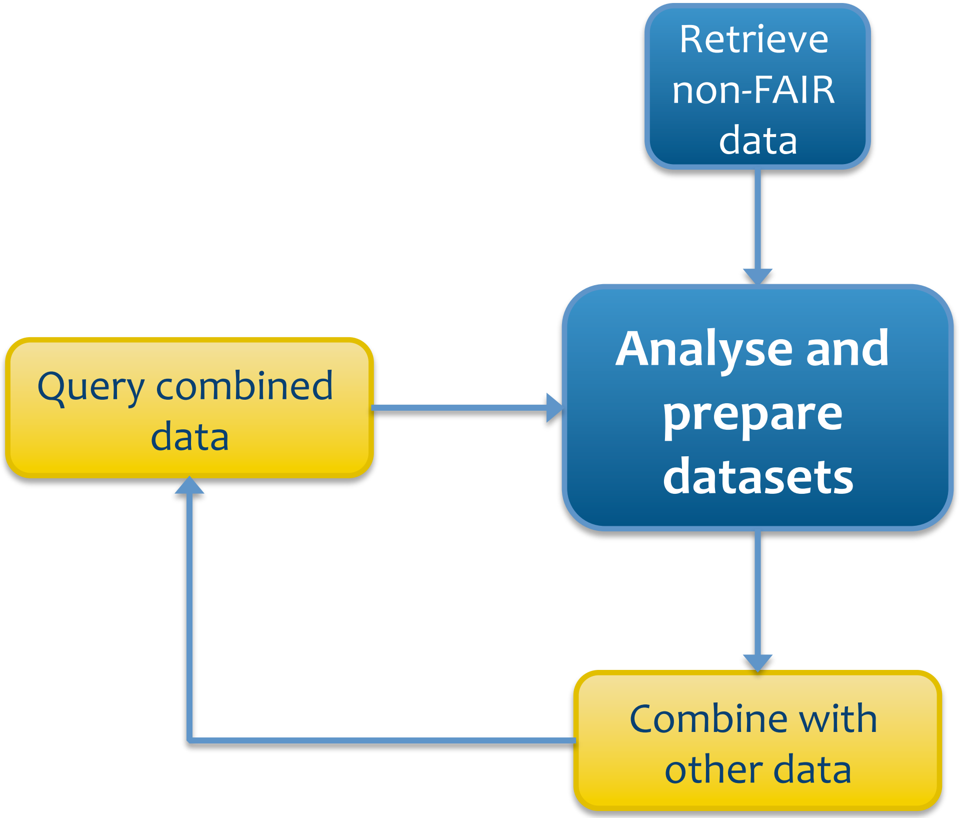
R1.2. (meta)data are associated with detailed provenance;

R1.3. (meta)data meet domain-relevant community standards;

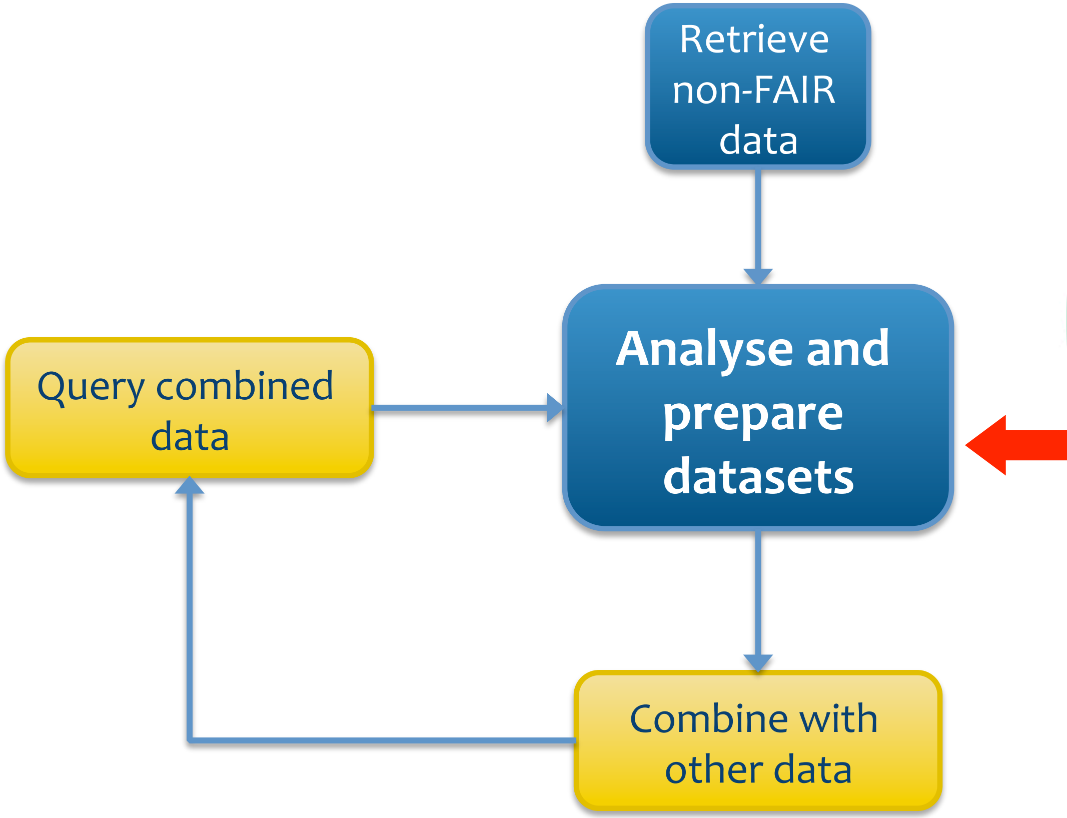
- **What do we need to fulfill this principle?**
  - Inform the digital resource's provenance information:
    - What are the core provenance information?
    - What are the community-specific provenance information?
    - How to represent provenance? Which vocabularies to use?
  - Inform the relevant community standards used by the digital resource (certification):
    - How to describe which standards are used?
    - How to describe compliance to these standards?
    - How to demonstrate that the standards are accepted by a given community?

FAIRIFICATION PROCESS

CURRENT SITUATION



CURRENT SITUATION

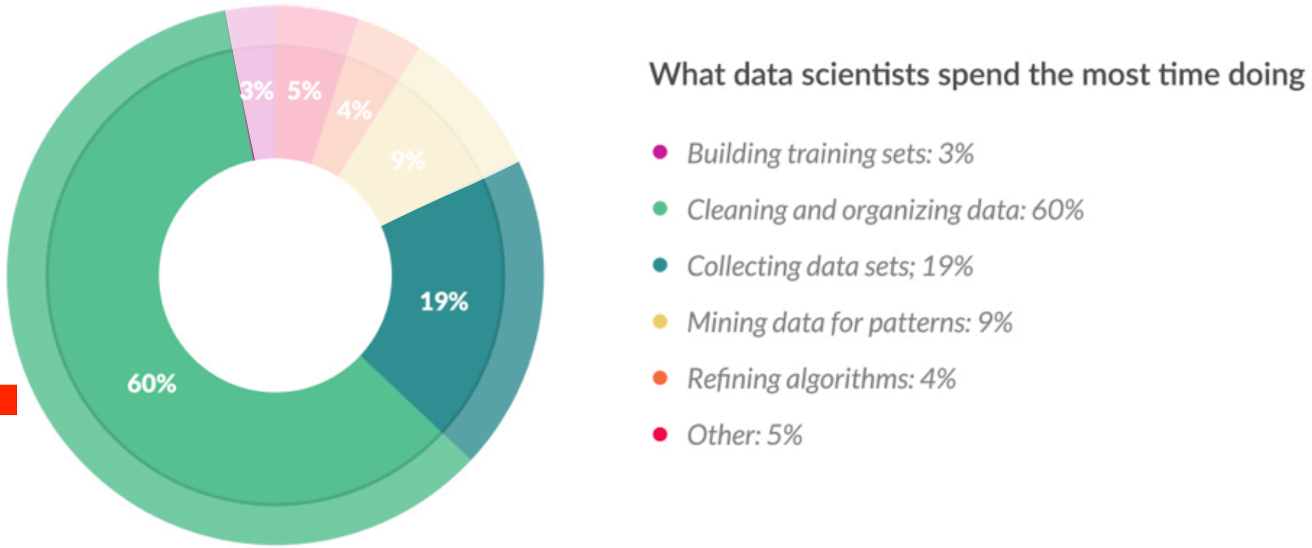


Retrieve non-FAIR data

Analyse and prepare datasets

Query combined data

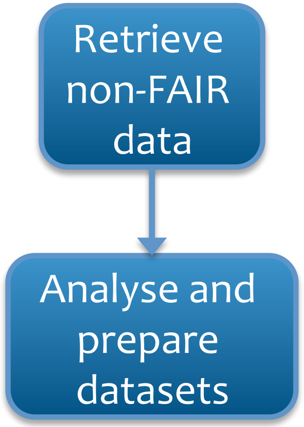
Combine with other data



FAIRIFICATION WORKFLOW

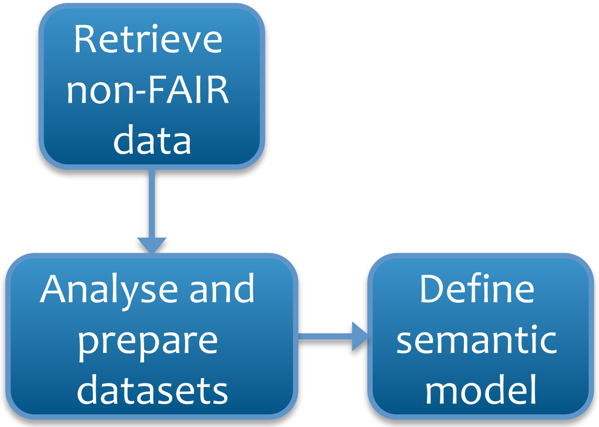
Download/locate file
Identify API call
Identify data access protocol

FAIRIFICATION WORKFLOW



Standard format (XML, RDF, relational DB API, VCF, DICOM, etc.)?
What is the content?
Column/field names?
Relations?
Data domain and range?
Understand the data
Data munging
…

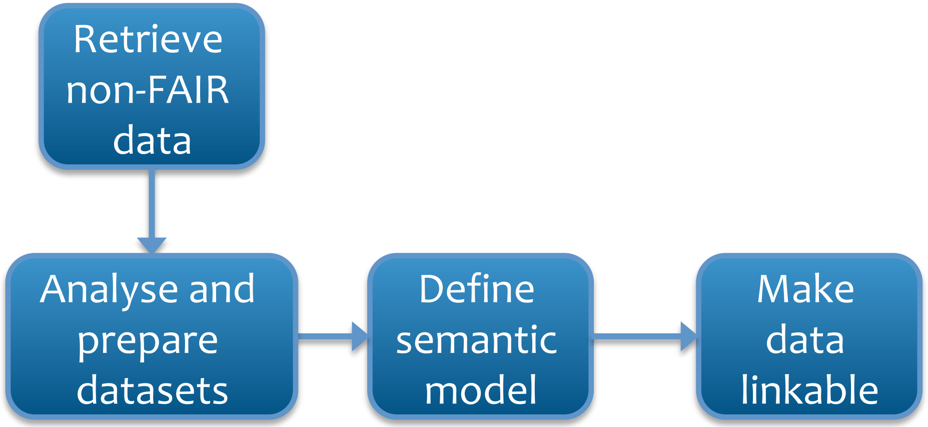
FAIRIFICATION WORKFLOW



What are the concepts involved?
Relations among concepts?
Existing vocabularies for the concepts and instances?

Interoperability
Reusability

FAIRIFICATION WORKFLOW

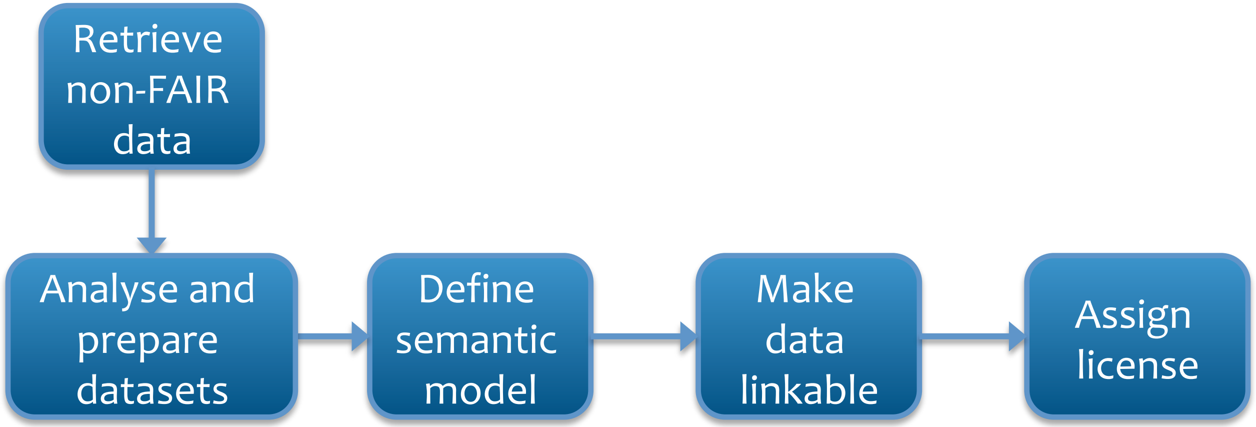


Apply the semantic model on the original data to make it linkable

Interoperability

Reusability

FAIRIFICATION WORKFLOW

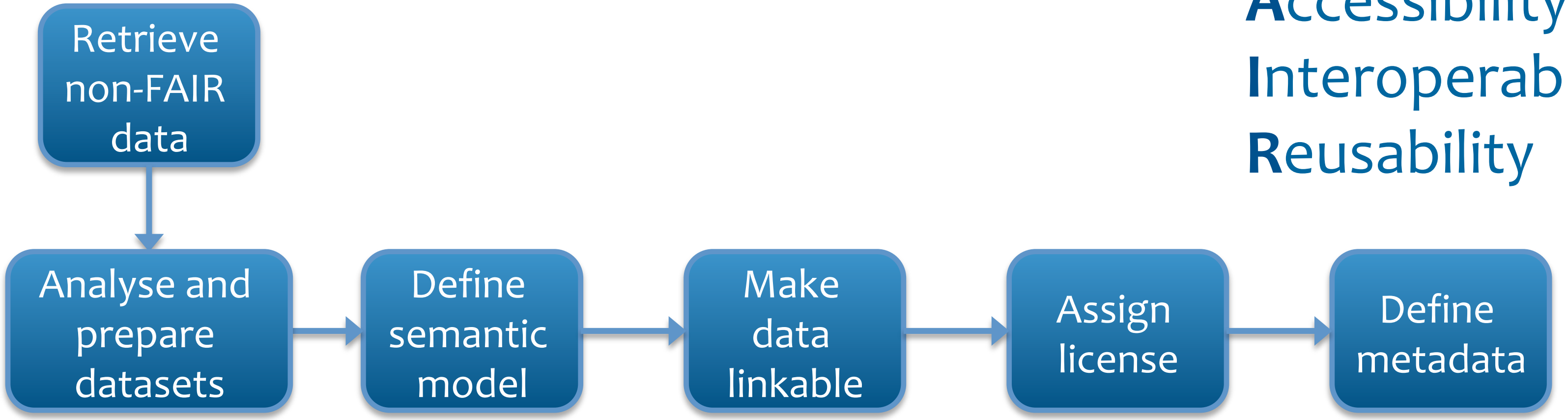


Who can access/reuse the data?
Under which conditions?

Accessibility
Reusability

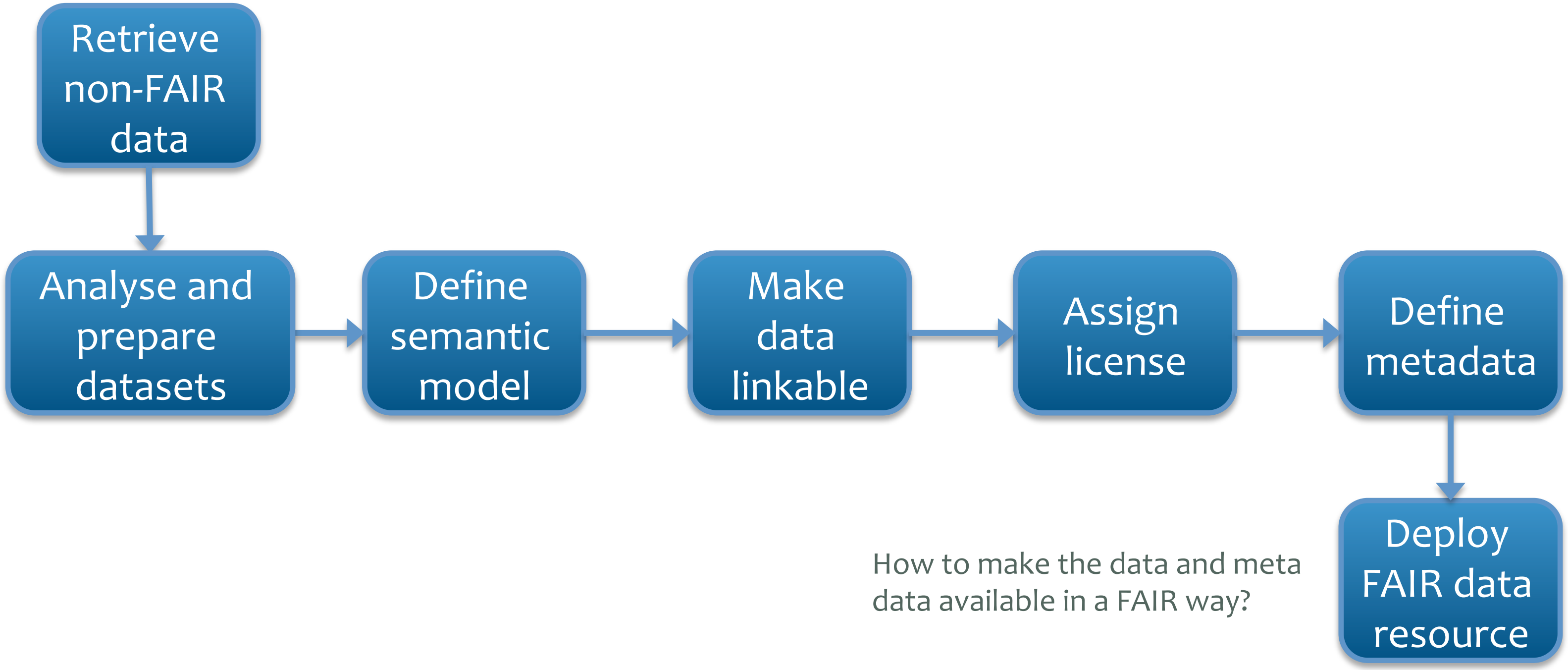
FAIRIFICATION WORKFLOW

Findability
Accessibility
Interoperability
Reusability



Authors
Version
Distributions
Provenance
…

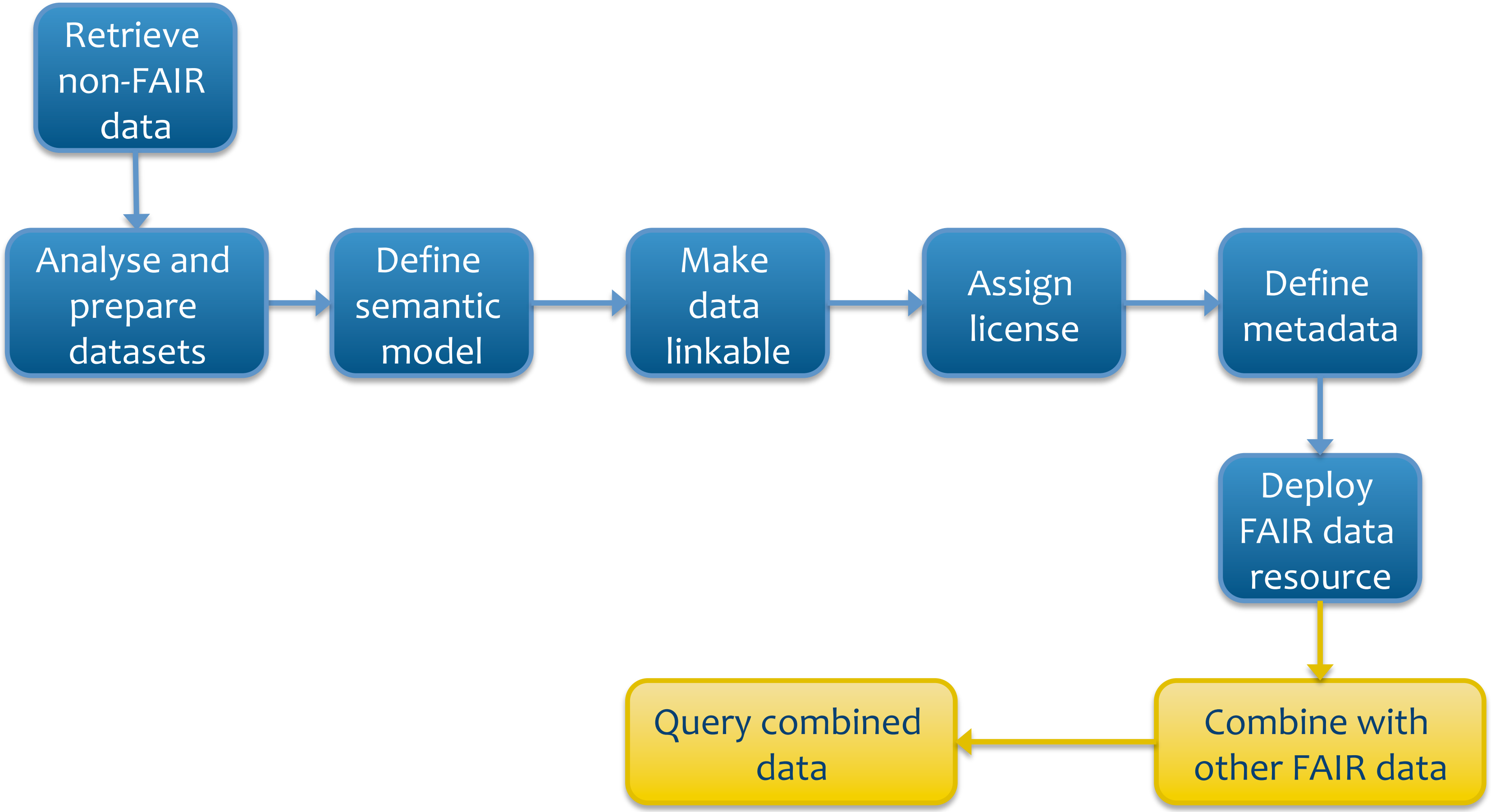
FAIRIFICATION WORKFLOW



How to make the data and meta data available in a FAIR way?

Findability
Accessibility
Interoperability
Reusability

FAIRIFICATION WORKFLOW



FAIRNESS ASSESSMENT CHALLENGES

WHY TO ASSESS?

- Because everybody is talking about FAIR and my resources should be seen as FAIR, whatever this means?

- To satisfy funders requirements?

- To serve as a guideline for achieving higher levels of interoperability and reuse with clarity on the concrete benefits (help improve)?

WHAT TO ASSESS?

- **Metadata and data?**

- **Only metadata?**

- **Only data?**
  - What do you mean by data?
  - In the FAIR principles, data refers to a variety of different resources, e.g., “traditional” data, services, software, APIs, vocabularies, ontologies, articles, etc.

HOW TO ASSESS?

- **Manual**
  - Takes advantage of human understandable artifacts, which are currently prevalent
  - May lead to subjective assessments and, therefore, harder to compare resources
  - Harder to scale
  - Harder to evaluate FAIR for machines, which is the main goal of the FAIR principles
- **Automatic**
  - Requires more rigor on the assessed resources
  - More likely to produce objective assessments
  - Easier to scale
  - Able to check if machines can, in fact, “work” with the (meta)data

HOW TO “READ” THE ASSESSMENTS?

- **Need for a scoring system**
  - One score for the 4 aspects of FAIR? Does not seem useful.
  - One score per aspect (F, A, I and R)?
  - One score per principle? What about the sub-principles?
  - Is there a hierarchy among the principles? Is there an order of precedence? Or different weights?
  - Is there an acceptable minimal FAIR level? Should it be across domains and applications or domain/community-dependent?
  - Do we use a pass/fail approach or introduce intermediary compliance levels in each/some evaluation?
- **Need for a visual representation of the scores**
  - To facilitate quick perception of the FAIRness level, a visual representation of the FAIR scores is required, e.g., stars, bars, etc.

GENERAL CHALLENGES

- Clarify that nobody has been asked to be 100% FAIR. Many times a lower FAIRness level is perfectly adequate.
- How to deal with the conflicting forces that, from one side want to push the communities towards a better (and FAIRer) data landscape and, from the other side, want to preserve the *status quo* (existing “kingdoms”) but labeling themselves FAIR?
- Who will define the assessment criteria?
- Who will execute the assessments based on the defined criteria?
- Should we have a unique set of assessment criteria? Or a core set for general comparison and domain-specific sets on top of the core for the specific needs of a given domain/application?

CURRENT STATUS OF THE FAIR METRICS

- Moving from metrics to maturity indicators
- The Maturity Indicator tests are also going to be "incremental". e.g. for the new I indicators there are "weak" and "strong" forms... with loose interpretation of "knowledge representation language" (e.g., CSV) vs strong interpretation (i.e. RDF)
- Full set of fully automatic evaluators almost complete
- Clear separation between the evaluation of metadata and data
- Used (together with the Data Stewardship Wizard) in the "FAIR Funders Pilot", involving Dutch ZonMW and Irish Health Research Board

OTHER CHALLENGES

- FAIR should be used as a guideline for achieving higher levels of interoperability and reuse with clarity on the concrete benefits.
- Improvements on FAIRness can be done incrementally, e.g., first deal with metadata then go for data.
- We need a combination of (FAIR) infrastructure and (FAIR) community practices.
- How to deal with other aspects not covered by the FAIR principles, e.g., openness, quality, … ?

Q&A – CONTACT INFO

GO FAIR

LUMC
Leiden University
Medical Center

Luiz Bonino

International Technology Coordinator – GO FAIR
Associate Professor BioSemantics – LUMC

E-mail: luiz.bonino@go-fair.org
Skype: **luizolavobonino**
Web: **www.go-fair.org**