FAIR-ification of LCA data using Semantic Web

AGNETA GHOSE

BONSAI

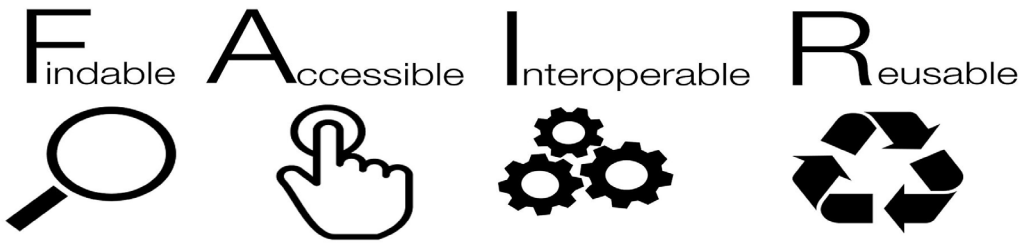
AALBORG UNIVERSITY
DENMARK

Challenge - Research data sits in different systems- hard to discover, obtain and integrate - eventually building new insight takes time

Why is it important to share research data?

- Makes research more open- transparency and reproducibility
- Makes your research more useful and citable by other researchers
- Allows data to be used to answer new questions
- Increasingly required by high impact journals
- Required by government funding agencies

FAIR guiding principles



Findable:

- Data are assigned with a unique identifier (e.g. doi).
- Registered and indexed in a searchable resource.
- Described with rich metadata.

Accessible:

- Data are retrievable by their identifier using a standardized protocol.
- Metadata should be accessible even if the data is proprietary.

Interoperable:

- Knowledge representation - Use a formal, accessible, shared and broadly acceptable language
- Data include qualified references to other data

Reusable:

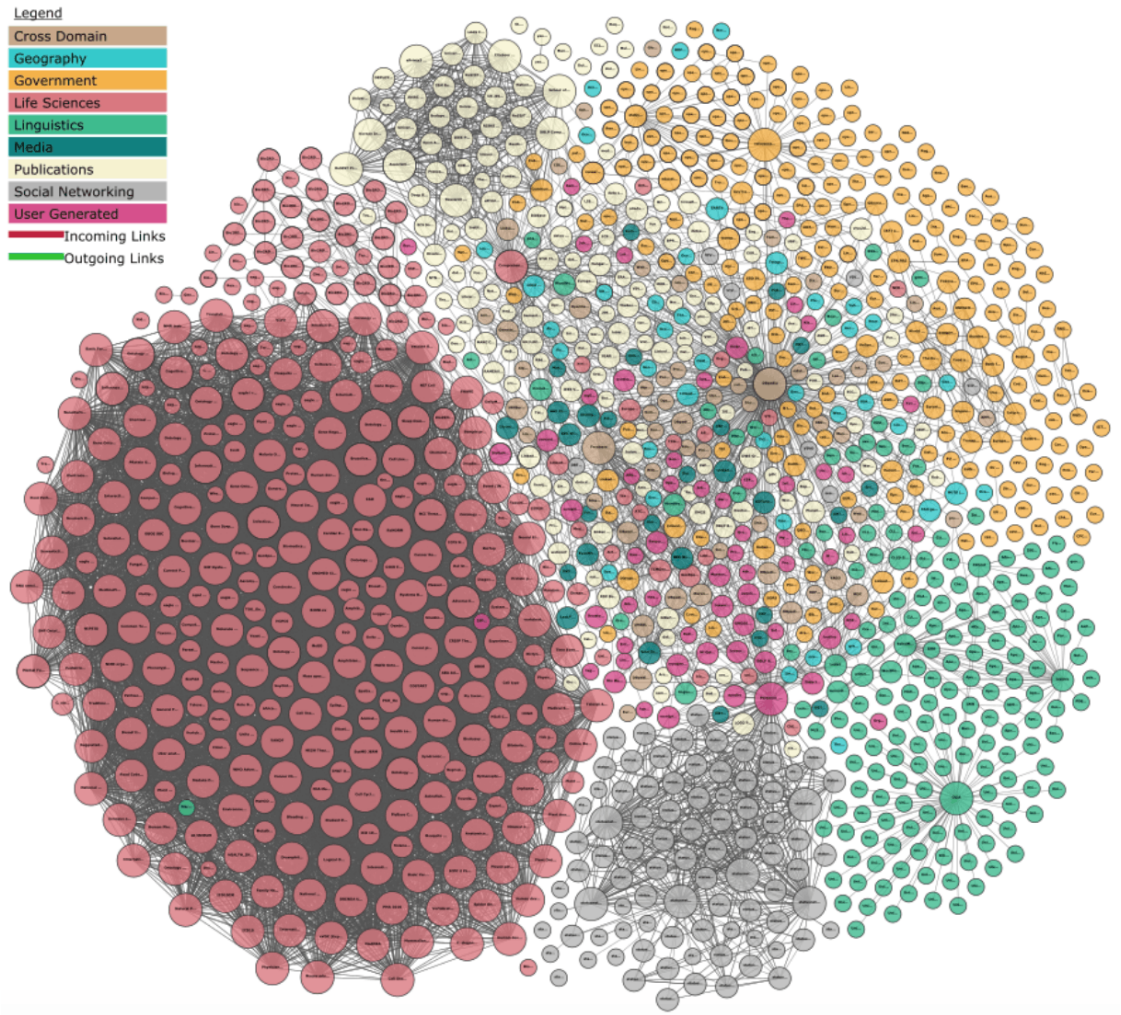
- Data are released with a clear, appropriate and accessible license.
- Data are associated with detailed provenance.

AALBORG UNIVERSITY
DENMARK

Source: Wilkinson's et.al
https://doi.org/10.1038/sdata.2016.18

Semantic Web and Linked Data

- World Wide Web – linked content/documents
- The key technology of the original web —was the hyperlink (from the end user's point of view)
- Semantic web is a cloud of linked data.
- Data gets an unique identifier
- Applications, that receive large amount of data from many different sources, benefit enormously from this feature.



Source: LOD-cloud.net

Semantic Web (contd.)

Semantic Web consists primarily of three technical standards:

1. **RDF (Resource Description Framework):** Data modeling language for the Semantic Web. Machine readable data format. All Semantic Web information is stored and represented in the RDF.

2. **OWL (Web Ontology Language):** The knowledge representation language of the Semantic Web. OWL enables to define concepts so that these concepts can be reused as much and as often as possible.

3. **SPARQL (SPARQL Protocol and RDF Query Language):** Query language of the Semantic Web. It is specifically designed to query data across various systems.

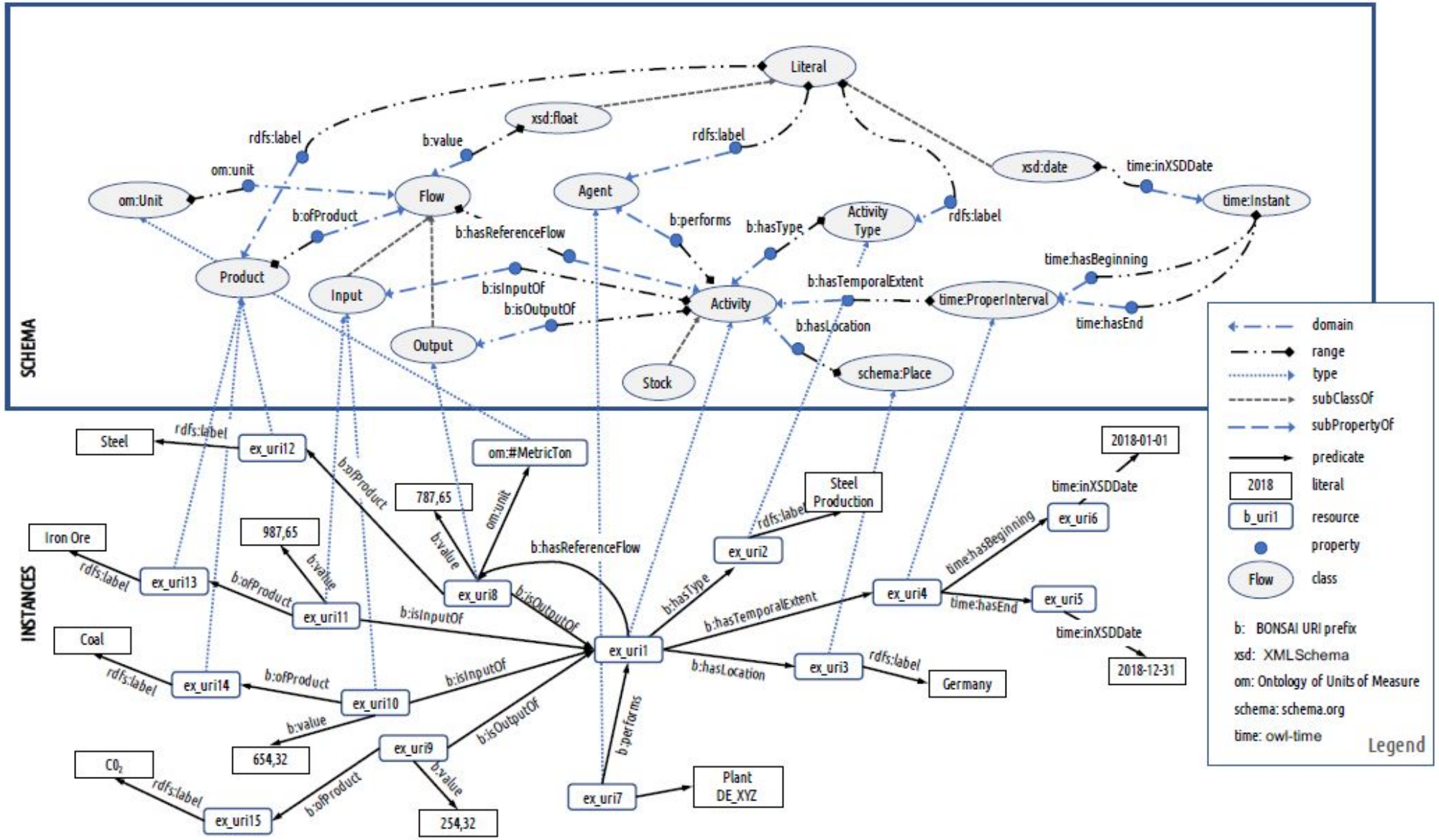
Why go semantic?

- To present knowledge about your data
- To allow data integration

What is the difference between RDF and OWL?

- Data formatted using these languages are machine readable and enhance automation.
- RDF is used to represent data in triple format, give it some structure and unique identifiers.
- OWL provides a rich vocabulary to allow reasoning and inference

Step 1 - Conceptualizing LCA on the Semantic Web – BONSAI ontology

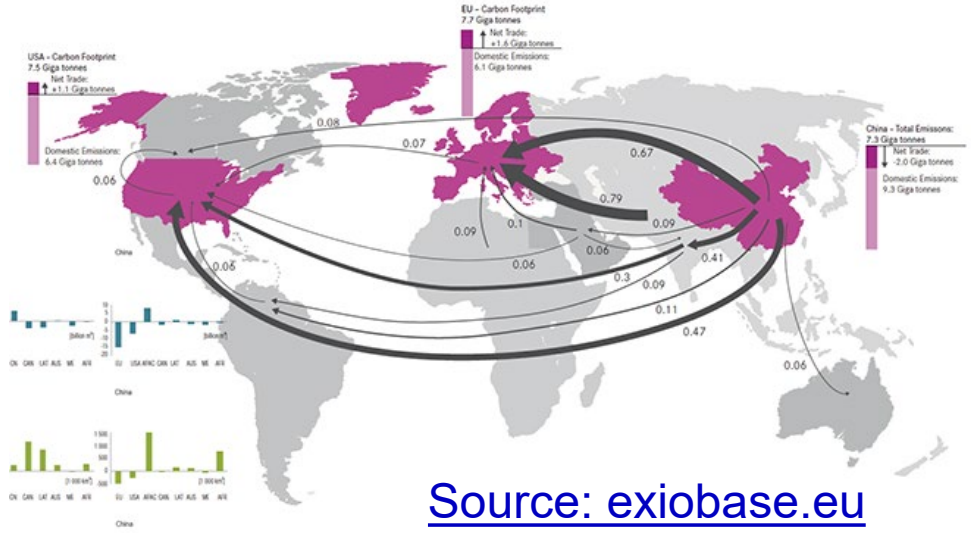


Step 2 – Convert existing open certified databases into machine readable RDF format

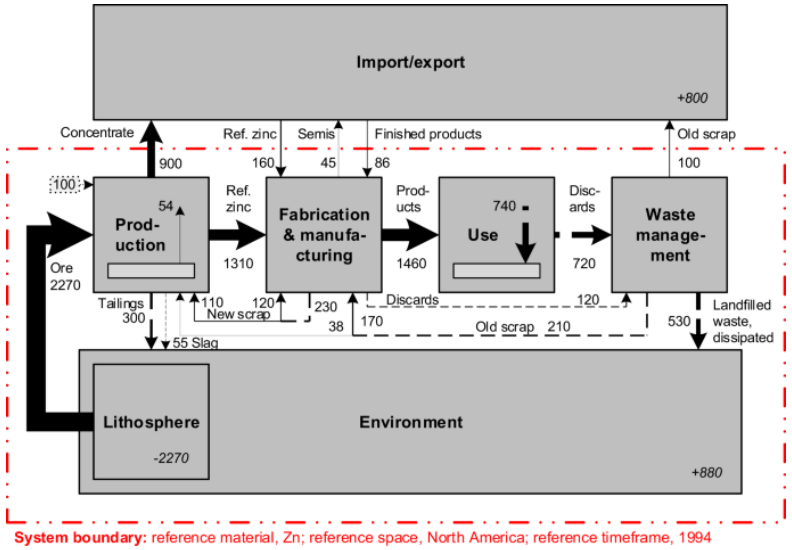
Two datasets:

1. Exiobase – multi-regional Input Output data base
2. Yale stocks and flow Database (YSTAFDB)

- Datasets developed using different modelling methods
  - Multi-regional Supply and Use Tables
  - Material Flow Analyis
- Different sources
  - Exiobase – specific website (needs sign up)
  - YSTAFDB – Zenodo.org
- Different data storage formats (csv, xslx)
- Certified as CC-BY 4.0



Source: exiobase.eu

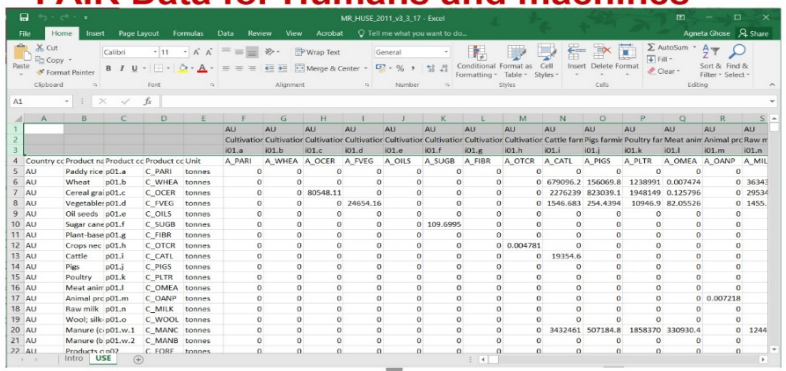


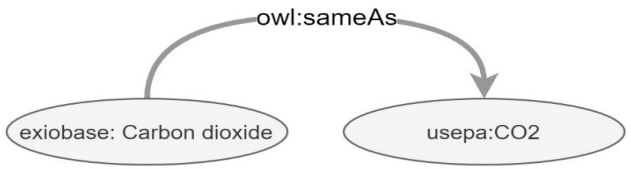
Source: Myers (2019)

AALBORG UNIVERSITY
DENMARK

Step 3: Develop SPARQL endpoint to query RDF data

FAIR Data for Humans and machines





Interoperability

Semantic mapping - gives meaning to data
Link ontology and data to existing ontologies and datasets

Step 1

Findability and Accessibility

Provides an unique identifier to each data instance

<http://rdf.bonsai.uno/flowobject/exiobase3_3_17#C_1> Carbon dioxide, fossil

Step 2

Re-Usability

Creation and querying of RDF Linked Data
Human and Machine understandable data

Step 3

RDF

Are semantic web the future for data sharing?

Opportunities

- Develop web applications and visualizations to automate the work process for LCSA researchers to integrate, query and explore data
- Use queried data directly with open LCA softwares
- Interoperability with major datasets such as data from FAOSTAT, Dbpedia, GIS sources etc can be a huge advantage in expanding the data

Challenges

- **Usability** - Researchers not accustomed to these models. Need for extensive programming knowledge
- **Infrastructure sustainability** – Setting up and maintaining servers. Project dependent
- **Data quality and validity**

THANK YOU!
BONSAI WIKI: GITHUB.COM/BONSAMURAIS/BONSAI/WIKI
GITHUB: GITHUB.COM/BONSAMURAIS
SPARQL ENDPOINT: HTTP://ODAS.AAU.DK/

AALBORG UNIVERSITY
DENMARK