

FAIR data principles and Metadata: or why build ontologies



Christopher Brewster (University of Maastricht):

Christopher.Brewster@maastrichtuniversity.nl



Funded by
the European Union

biofin-project.eu

Of metadata, of ontologies, and of FAIR data Principles



- Strange terms, strange concepts for most people
- Why are we interested in this in the BioFIN project?
- What is this?
- How do we do it?
- Why do we do this?
- ... but first a little story ... almost a history lesson

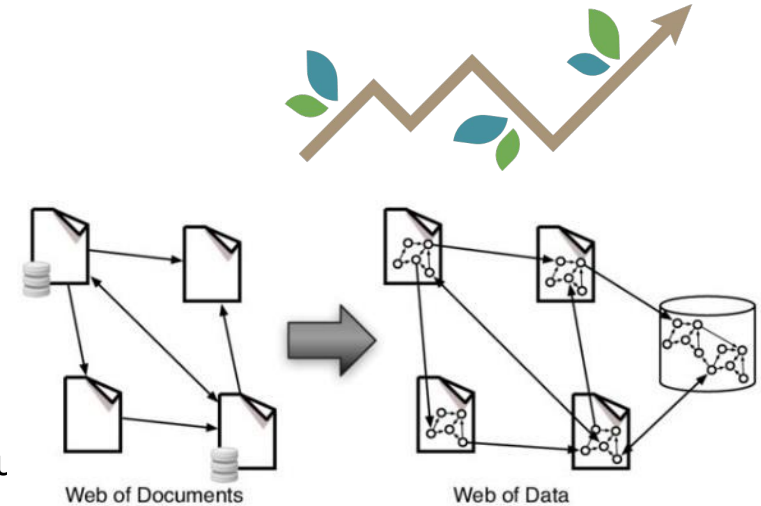
The time has come,' the Walrus said,
To talk of many things:
Of shoes — and ships — and sealing-wax —
Of cabbages — and kings —
And why the sea is boiling hot —
And whether pigs have wings.'

-- Lewis Carroll



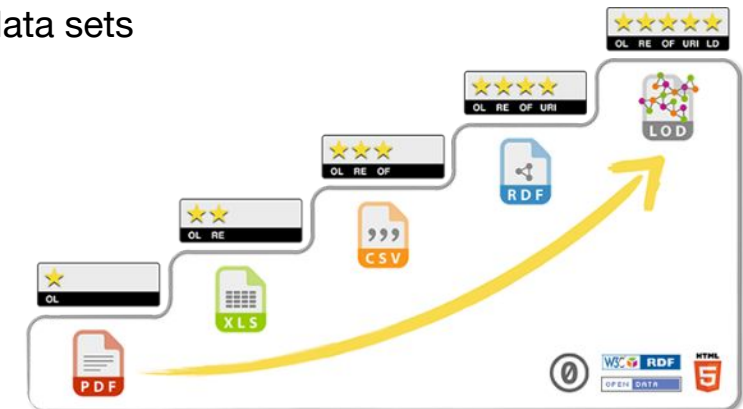
Story 1 – The Web of Data

- 1992 Tim Berners-Lee invents the World-Wide-Web. This was designed as a web of documents.
- TBL realises that a web of documents was insufficient and what was needed was a “web of data”.
- From this realization arose a series of technologies we generally call "**semantic web**" - to gradually transform the web into a web of data.
 - Includes standards such RDF, RDFS, OWL, SPARQL, and lots more under the aegis of W3C



- In 00s, TBL proposed the idea of "Linked Data“:

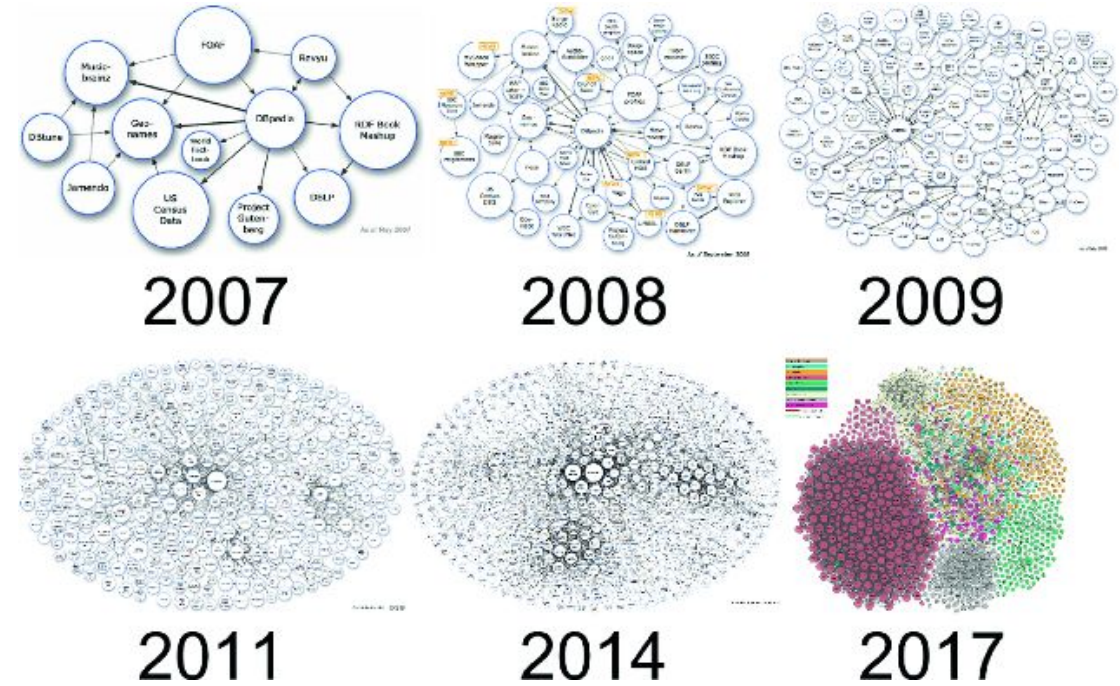
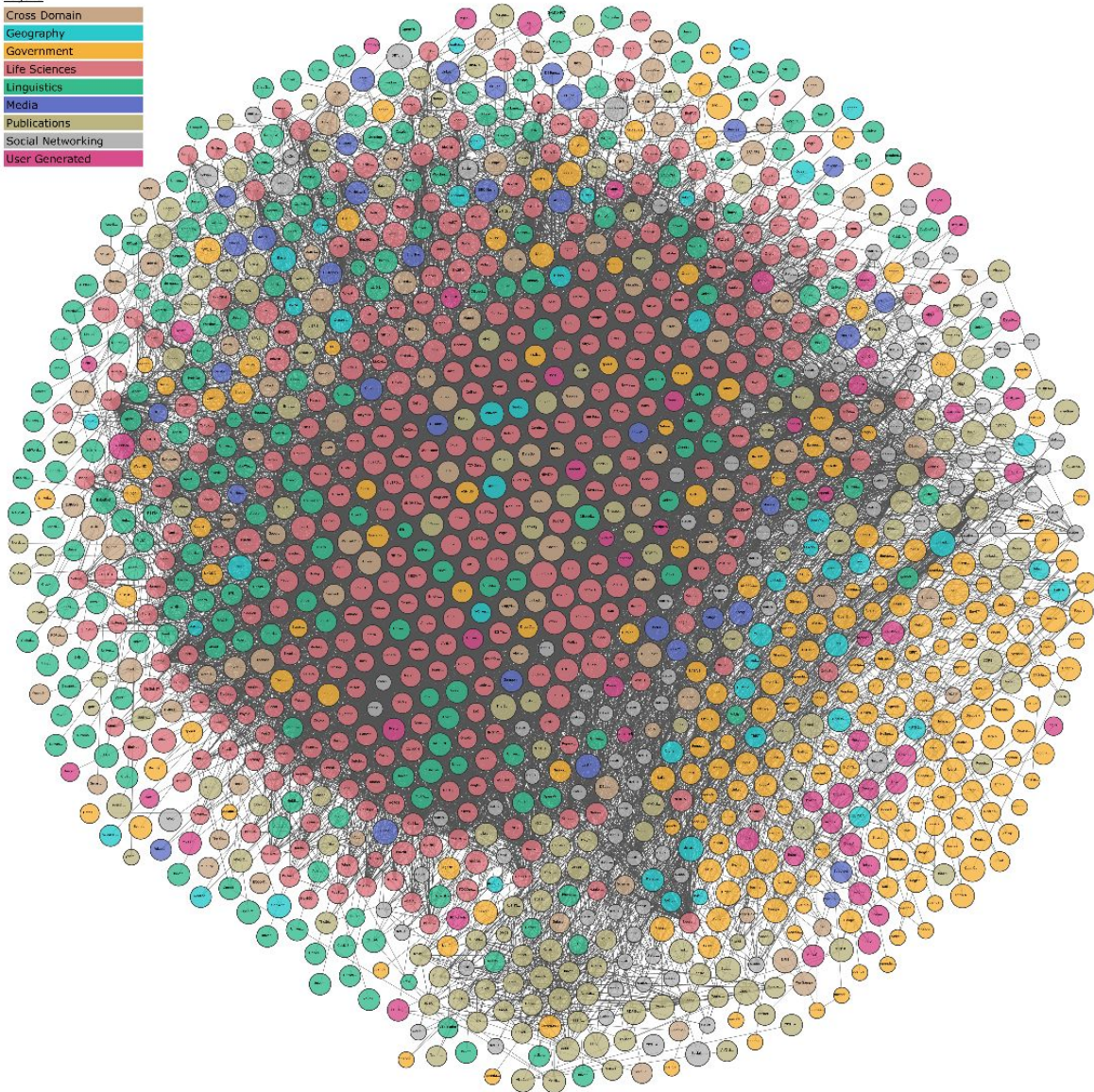
- One ★ for online in any format – “open data”
- Two ★★ for online in machine readable format e.g Excel
- Three ★★★ for online, in non-proprietary format e.g. csv
- Four ★★★★ for online, non-proprietary format, use open standards to identify stuff (i.e. use URIs, RDF etc.)
- Five ★★★★★ for online, non-proprietary format, use open standards, link to other data sets



Story 2 Linked Open Data Cloud



- Legend
- Cross Domain
 - Geography
 - Government
 - Life Sciences
 - Linguistics
 - Media
 - Publications
 - Social Networking
 - User Generated



Story 3 Open Science

- Two contrary movements
 - Panic about "open data", problem especially in health but general move towards respecting privacy, ownership etc.
 - Frustration with research being paid for but not open, accessible, frustration both from scientists and funding agencies, some parts of general public/politicians
 - Frustration also research gets lost, inaccessible, loss of context etc.
- Result (cutting a long story short)
 - European Open Science Cloud - from the EC
 - FAIR Data Principles - from the Life Science community

<https://unsplash.com/photos/PdDBTrk>



The FAIR Data Principles



- Important paper laid the foundations: Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Has had a huge impact ... generally adopted by the EC and many other funding agencies
- What does it mean?



Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be 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

To be 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


To be 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 Consequences




The **FAIR** data principles

 **F**indable
To identify data for both humans and computers by computerising metadata that facilitate searching for specific datasets.


 **A**ccessible
Data is stored properly -for long term- so that it can easily be accessed and/or downloaded with well-defined access conditions. These could be access to the metadata (only) or getting access to the actual data.

 **I**nteroperable
The ability to combine different datasets either by humans or by computers. Therefore multiple agreements have to be made with respect to the terminology used to prevent ambiguities of the meanings of these terms.


 **R**eusable
Data should be ready to be used for future research and to be further processed using computational methods. This requires adequate information about how the data were obtained and processed (provenance), and an appropriate license.

<https://www.dtls.nl/fair-data/data-stewardship/>

Steps how to make data **FAIR**

 **F**indable


- select a data repository at an early stage and check out its data format and metadata requirements
- make sure the data can get a persistent identifier so that it can be cited
- select a catalogue to make your data more findable, especially if the repository is more generic in nature

 **A**ccessible

- guarantee longevity of the data (i.e., by submitting it to a repository that has a certification like e.g. ISO)
- check and describe the legal conditions under which the data can be made available
- establish an embargo period if necessary
- make sure your ICT infrastructure will keep the data available even in case of equipment failure or human error

 **I**nteroperable

- select commonly used data formats
- select commonly used vocabularies for data items

 **R**eusable

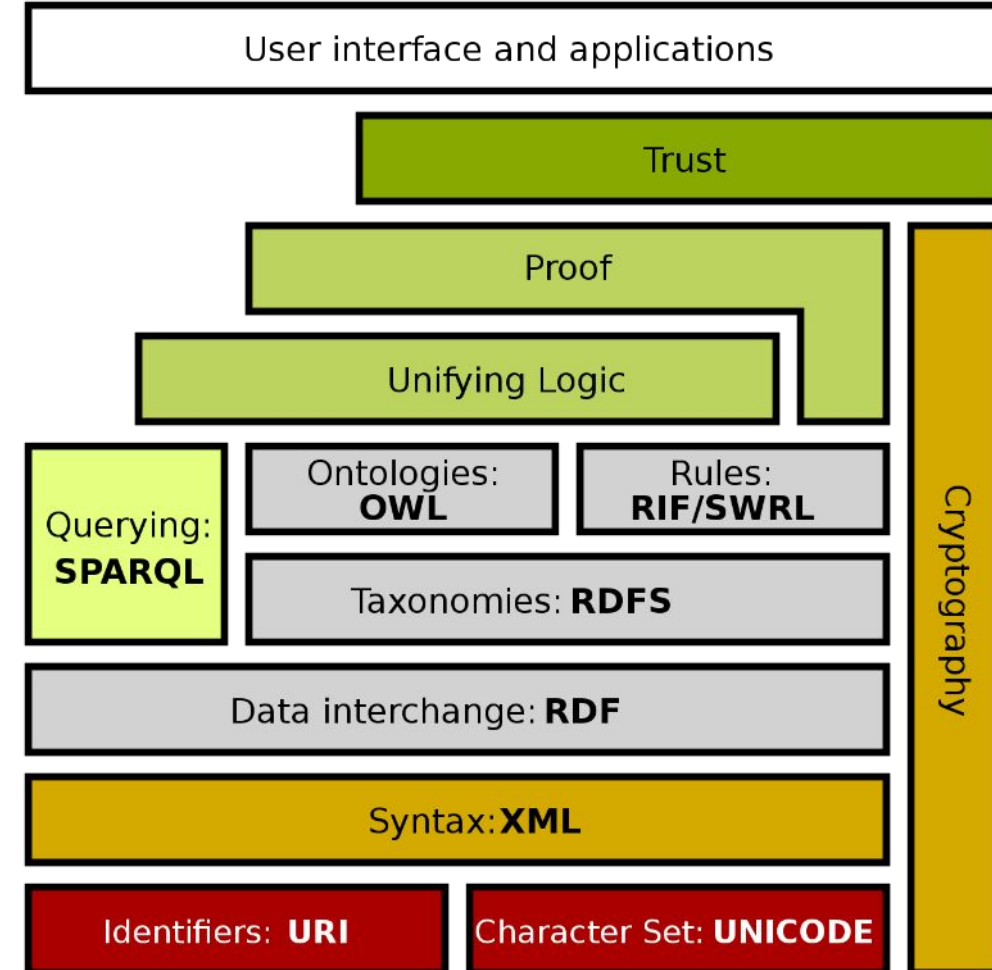
- make sure you keep proper provenance information (i.e., details about how and where the data was generated, incl. machine settings, details about processing steps: the software tools with their versions and parameters)
- select the right minimal metadata standard and collect the necessary metadata (many minimal metadata standards are included in ELIXIR's biosharing.org repository)
- select a license for the data (preferably an open license) and the associated software tools
- make sure the important conclusions of your study will not only be available in a paper in a narrated form, but also in a digital file (e.g., a nanopublication)

<https://www.dtls.nl/fair-data/fair-data-knowledge-expertise/>

FAIR Consequences – semantic technologies



- Metadata for a NbS to be findable
 - Need to use an ontology/taxonomy/vocabulary that is widely used to label the NBS/data with appropriate keywords/concepts
 - Need have unique identifiers
- Metadata for NBS to be accessible
 - Need to use a commonly used protocol to access the NbS/data
 - Need to have access controls – who is allowed to have access to that data?
- Metadata for aNbS to be interoperable
 - Need to agreed ontology to describe the NbS/data, even more important if data is to be machine readable
 - Ontology must be following FAIR principles as well
- Metadata for a NbS to be reusable
 - Need for provenance data – where did this NbS come? Who made it?
 - Need for suitable machine readable licences



Example Metadata



Zenodo – some article - DC format

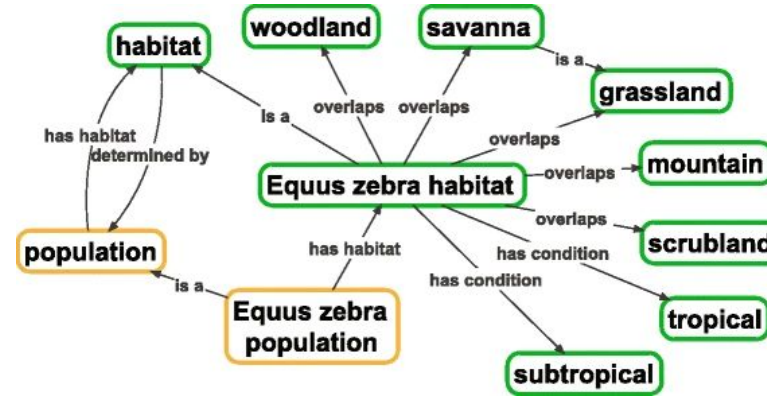
```
<?xml version='1.0' encoding='utf-8'?>
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:creator>Marutsov, Plamen Dimitrov</dc:creator>
  <dc:date>2014-01-14</dc:date>
  <dc:description>Mycotoxins are toxic compounds (secondary metabolites) produced by various saprophytic living mold fungi belonging to genera Aspergillus, Fusarium, Penicillium, Claviceps, Alternaria, and others. They are formed and accumulated as a result from proliferation of molds on a variety of food substrates under favorable environmental conditions, including a suitable temperature and humidity. The term 'mycotoxin' is a combination from the Greek word mykos - fungus, mold, and the Latin word 'toxicum' - poison. For the first time, the term mycotoxins was used in England in 1960 after detecting of high mortality in young turkeys in a turkey farm close to London ('Turkey-X disease'). After the tests that were carried out, high contents of aflatoxins were found out in the peanut butter originating from Brazil that was added to the feed. (Blount, W. P. 1961, Allcroft et al., 1961). By now, the number of the mycotoxins known is over 400, and generally are identified more than 30 000 different metabolites produced by molds.</dc:description>
  <dc:description>BG; en; EFSAfocalpoint@mzh.government.bg</dc:description>
  <dc:identifier>https://zenodo.org/record/826599</dc:identifier>
  <dc:identifier>10.5281/zenodo.826599</dc:identifier>
  <dc:identifier>oai:zenodo.org:826599</dc:identifier>
  <dc:relation>doi:10.5281/zenodo.826598</dc:relation>
  <dc:relation>url:https://zenodo.org/communities/efsa-kj</dc:relation>
  <dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
  <dc:rights>https://creativecommons.org/licenses/by/4.0/legalcode</dc:rights>
  <dc:subject>Bulgaria</dc:subject>
  <dc:subject>Opinion</dc:subject>
  <dc:subject>mycotoxins</dc:subject>
  <dc:subject>molds</dc:subject>
  <dc:subject>agriculture</dc:subject>
  <dc:subject>mycotoxins</dc:subject>
  <dc:subject>molds</dc:subject>
  <dc:subject>agriculture</dc:subject>
  <dc:title>Epidemiological and social aspects of mycotoxins in dairy agriculture</dc:title>
  <dc:type>info:eu-repo/semantics/report</dc:type>
  <dc:type>publication-report</dc:type>
</oai_dc:dc>
```

Wikidata – Maastricht – in RDF

```
<rdf:RDF>
<rdf:Description rdf:about="https://www.wikidata.org/wiki/Special:EntityData/Q1309">
<rdf:type rdf:resource="http://schema.org/Dataset"/>
<schema:about rdf:resource="http://www.wikidata.org/entity/Q1309"/>
<cc:license rdf:resource="http://creativecommons.org/publicdomain/zero/1.0"/>
<schema:softwareVersion>1.0.0</schema:softwareVersion>
<schema:version
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1831142180</schema:version>
<schema:dateModified
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2023-02-10T18:20:04Z</schema:dateModified>
<wikibase:statements
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">178</wikibase:statements>
<wikibase:sitelinks
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">121</wikibase:sitelinks>
<wikibase:identifiers
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">51</wikibase:identifiers>
</rdf:Description>
<rdf:Description rdf:about="http://www.wikidata.org/entity/Q1309">
<rdf:type rdf:resource="http://wikiba.se/ontology#Item"/>
</rdf:Description>
<rdf:Description rdf:about="https://af.wikipedia.org/wiki/Maastricht">
<rdf:type rdf:resource="http://schema.org/Article"/>
<schema:about rdf:resource="http://www.wikidata.org/entity/Q1309"/>
```

Ontologies

- What is an ontology? Just a machine readable, formal way of describing a part of the world.
- There are lost of ontologies Central to Linked Data, central to any form of “knowledge representation”
- Typically use RDF/RDFS/OWL formalisms to be machine readable
- Lots and lots of agriculture, forestry and environment ontologies e.g. AGROVOC, FOODON, AGRO, ENVO
 - Too many, often lack of agreement means every organisation goes and creates another one ...
 - However, necessary to achieve interoperability



ENVO

YSO - General Finnish ontology

Content language English

A-Z Hierarchy Groups New and Deprecated

objects > place > areas and regions > areas created by nature > forests

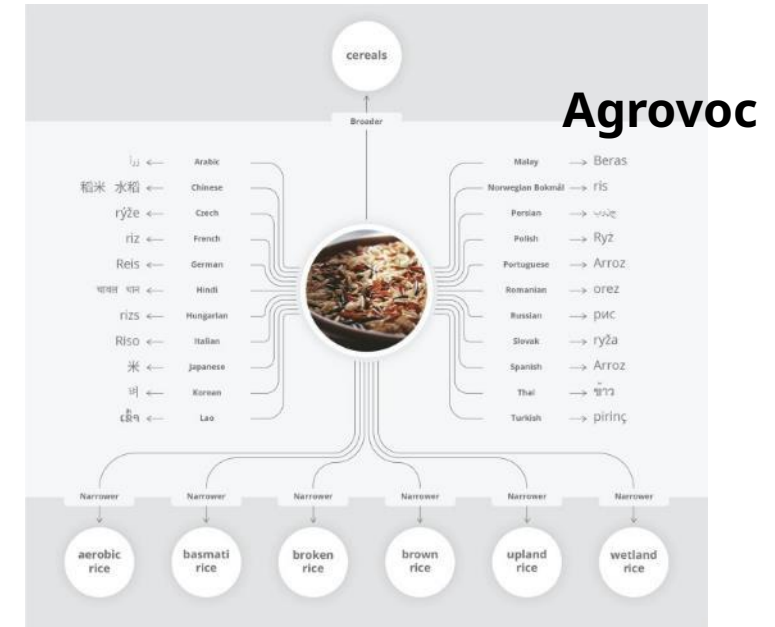
PREFERRED TERM **forests**

TYPE General concept

BROADER CONCEPT areas created by nature

NARROWER CONCEPTS

- forests by age
 - old growth forests
 - primeval forests
 - young stands
- forests by location
 - Finnskog areas
 - hilltop forests
 - urban forests
- forests by management
 - cultivated forests
 - natural forests
 - primeval forests
- forests by ownership
 - jointly owned forests



<http://finto.fi/ysso/en/page/p5454>

