Ontology for food and nutrition: a solution for multiple level measurement, data integration and analysis

Chen Yang
Ghent University
Chen.Yang@UGent.be
Get scientific conclusion?

small-scale
Specific

big-scale
General

small-scale
Specific

Too many hypothesis

Too hard

When???
Multiple-level data collection: easy?

Sample with enough power

• Money?
• Time?
• Staff?
• Follow-up?
Systematic review: easy?

• Study?
• Data?
• Next challenges:
  • Open access?
  • Re-usable?

**Supplementary material 1**: Search syntax for systematic literature review of existing instruments for quality of observational studies

<table>
<thead>
<tr>
<th>Search Database</th>
<th>PubMed/Medline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search date</td>
<td>July 1st 2015</td>
</tr>
</tbody>
</table>

Search syntax:

Search (((((((Tool[Title/Abstract]) OR Score[Title/Abstract]) OR Scale[Title/Abstract]) OR Instrument[Title/Abstract]) OR Checklist[Title/Abstract]) OR Item[Title/Abstract]) AND ((("quality score"[Title/Abstract]) OR quality assessment tool*[Title/Abstract]) OR "Critical appraisal"[Title/Abstract]) OR "Methodological quality"[Title/Abstract]) OR "Methodological quality"[Title/Abstract]) OR Evidence-Based Medicine/methods[MeSH Terms]) OR "Methodological quality"[Title/Abstract]) OR Evidence-Based Medicine/methods[MeSH Terms])

Schema: nomsch Filters: Publication date from 2000/01/01 to 2015/12/31; English
Get scientific conclusion?

small-scale  big-scale  small-scale
Specific     General     Specific

Too hard
Open Access and FAIR principle

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplary implementations in the community.

Supporting discovery through good data management
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to advancement data and knowledge integration and reuse for the
FAIR principle
(Mark D. Wilkinson et al. 2016)

• Findable
• Accessible
• Interoperable
• Reusable
FAIR principle
(Mark D. Wilkinson et al. 2016)

- Findable
- Accessible
- Interoperable
- Reusable

Standardized data descriptions
Machine-readable
Different data descriptions

Search: “Dietary intake dataset”

- Dietary intake dataset
- Food and nutrition dataset
- 营养学数据
The Gap: “Messy” data management

• Different data formats (excel file, csv,...)
• Different data terminologies
• Different languages
• Different data qualities
• Different units for data
• Lack of machine-readable description
Result?

- Different data formats (excel file, csv, ...)
- Different data terminologies
- Different languages
- Different data qualities
- Different units for data
- Lack of machine-readable descriptions

FAIR principle
Data integration

Data loss
Simple solution: Unified (meta-) data identifier?

Search: “Dietary intake data”
Benefits of (meta-) data identifier

• Harmonized data formats
• Harmonized data terminologies
• Harmonized languages
• Harmonized units for data
• Harmonized well-described data qualities
• machine-readable language

✓ Standardized storage
✓ Findable
✓ Accessible
✓ Interoperable
✓ Re-usable
✓ Automatic Harmonization
Benefits

- Substantial existing data resources
- Adequately power
- Increase return on research investments
- More research possibilities with data comparisons from difference resources

Make “Multiple-level big data analysis” easier?

General conclusion?
Standardized nutrition dictionary?

• Machine-readable **No**
• For searchable databases **No**
What is ontology? (Noy & McGuinness, 2001)

- A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.
- Human-readable
- **Machine-readable (computer language)**
- **Structure the group of concepts**

![Diagram showing the structure of a dataset related to ENPADASI project with categories such as 'Participants features', 'Data collection methods', 'Study types', and 'Dataset China'.]
Using concepts to describe a research

The dataset China
✓ recruited participants who are...
✓ collected data by using the methods...
✓ is for study type...
✓ ...

Identifiers

Dataset China

- Participants features
- Data collection methods
- Study types
- ...

...
To share information among datasets

• (Automatic) Extract information from different sites
• (Automatic) Aggregate information from different sites
ontology links different levels?

Find similar data at different level for stronger conclusion?

Figure 1. Food and nutrition security distal, intermediate and proximal determinants. Conceptual framework adapted with slight modifications from Smith.

Note: Frankenberg et al.

Pérez-Escamilla & Segall-Corrêa (2008)
To enable reuse of domain knowledge

• Re-use ontology

Existing data

<table>
<thead>
<tr>
<th>Year</th>
<th>Data EU</th>
<th>Ontology for EU2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Future data

<table>
<thead>
<tr>
<th>Year</th>
<th>Data EU</th>
<th>Ontology for EU2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Current work**

<table>
<thead>
<tr>
<th>No.</th>
<th>Dietary record</th>
<th>24-h recall</th>
<th>FFQ and screener</th>
<th>Dietary history</th>
<th>Descriptors</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Type of administration</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Origin of the questionnaire</td>
</tr>
</tbody>
</table>

- Proxy administered
- Self-administered and not verified by interviewer
- Self-administered and checked by interviewer
- Interview administered
- Interview administered using automated multiple-pass method
- Self-developed questionnaires
- Use of standardized questionnaire
- Adopted other questionnaires

---

**Perspective: Essential Study Quality Descriptors for Data from Nutritional Epidemiologic Research**

Potential framework

- **Data annotation**
  - term (FoodEx2), nutrients (Food composition table)

- **Meta data annotation**
  - Minimal data requirement

- **Data quality description**

- **Study description (reporting guidelines: STROBE-nut)**
  - Methods, result, etc.
Findable data WITHOUT searching studies?

- Require data
- Search studies
- Search data

[Diagram showing the process with 'Search studies' crossed out]
Platform

• DISQOVER (based on ontology)
• Solution for food and nutrition measurement?
Future plans

• Describe every detail of Food and nutrition study and data

• Cooperation