How FAIR am I?
FAIR Principles and Interoperability of Data and Tools

Peter Doorn, DANS
@pkdoorn @dansknaw

Plan-Europe - Platform of National eScience Centers in Europe
PLAN-E meeting, April 27 & 28, 2017, Poznan, PSNC, Poland

Acknowledgments: Ingrid Dillo (DSA) and Emily Thomas (FAIR Data Assessment Tool)
DANS is about keeping data FAIR

Mission: promote and provide permanent access to digital research resources

Institute of Dutch Academy and Research Funding Organisation (KNAW & NWO) since 2005

First predecessor dates back to 1964 (Steinmetz Foundation), Historical Data Archive 1989
What is Interoperability?

**Interoperability** (pronounced *IHN-tuhr-AHP-uhr-uh-BIHL-ih-tee 😊*) is the ability of a system or a product to work with other systems or products without special effort on the part of the user.

Source: various dictionaries
Interoperability is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, in either implementation or access, without any restrictions.

Degrees of interoperability:
Interoperable data definitions

• Interoperability describes the extent to which systems and devices can exchange data, and interpret that shared data, without any restriction on access and implementation of the data.

• Data interoperability reflects our ability to let computers find, access and utilise data from physically separate and heterogeneous data repositories.
  • (Condition: must be machine readable = the ability for a computer to extract a description of the terms and conditions from a licence document in order to compare and combine to similar data sets)

Note that Interoperability is seen as the resultant of Findability, Accessibility and Usability in the last definition.
Three “Levels” of Data Interoperability

• **Foundational** interoperability is establishing the basic ability for two or more systems to *exchange* data. This level allows data exchange from one IT system to be received by another (and does not require the ability for the receiving IT system to interpret the data).

• **Structural** interoperability defines the *syntax* of the data exchange. It ensures that data exchanges between IT systems can be interpreted at the data field level.

• **Semantic** interoperability is the ability for two or more systems to effectively exchange, *interpret* and *use* data and information.

Source: [https://swc.net/general/blogs/what-interoperability](https://swc.net/general/blogs/what-interoperability)
DANS and Data Seal of Approval (DSA)

- 2005: DANS to promote and provide permanent access to digital research resources

- Formulate quality guidelines for digital repositories, including DANS

- 2006: **5 basic principles** as basis for 16 DSA guidelines

- 2009: international DSA Board

- Almost 70 seals acquired around the globe, but with a focus on Europe
The Certification Pyramid

ISO 16363:2012 - Audit and certification of trustworthy digital repositories
http://www.iso16363.org/

DIN 31644 standard “Criteria for trustworthy digital archives”
http://www.langzeitarchivierung.de

http://www.datasealofapproval.org/
https://www.icsu-wds.org/
New common requirements for data repositories by DSA and World Data system (WDS)

18 Requirements:

• Context (1)
• Organizational infrastructure (6)
• Digital object management (8)
• Technology (2)
• Additional information and applicant feedback (1)

https://goo.gl/kZb1Ga

Endorsed recommendation by Research Data Alliance
European Commission Recognition
Resemblance DSA – FAIR principles

<table>
<thead>
<tr>
<th>DSA Principles (for data repositories)</th>
<th>FAIR Principles (for data sets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>data can be <strong>found</strong> on the internet</td>
<td>Findable</td>
</tr>
<tr>
<td>data are <strong>accessible</strong></td>
<td>Accessible</td>
</tr>
<tr>
<td>data are in a <strong>usable format</strong></td>
<td>Interoperable</td>
</tr>
<tr>
<td>data are <strong>reliable</strong></td>
<td>Reusable</td>
</tr>
<tr>
<td>data can be <strong>referred</strong> to</td>
<td>(citable)</td>
</tr>
</tbody>
</table>

The resemblance is not perfect:
- “usable format” (DSA) is an aspect of “interoperable” (FAIR)
- FAIR explicitly addresses machine readability

A certified TDR already offers a baseline data quality level
Combine and operationalize: DSA & FAIR

- Growing demand for quality criteria for research datasets and a way to assess their fitness for use

- Combine the principles of core repository certification and FAIR

- Use the principles as quality criteria:
  - Core certification – digital repositories
  - FAIR – research data (sets)

- Operationalize the principles as an instrument to assess FAIRness of existing datasets in certified TDRs
All data sets in a Trusted Repository are FAIR, but some are more FAIR than others.
Implementing the FAIR Principles

To be Findable:
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

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 Re-usable:
R1. meta(data) have 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 their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Badges for assessing aspects of data quality and “openness”

These badges do not define good practice, they certify that a particular practice was followed.

BRONZE: data is openly licensed, available with no restrictions, accessible and legally reusable.

SILVER: satisfies the Bronze requirements, the data is documented in a machine readable format, reliable and offers ongoing support from the publisher via a dedicated communication channel.

GOLD: satisfies the Silver requirements, is published in an open standard machine readable format, has guaranteed regular updates, offers greater support, documentation, and includes a machine readable rights statement.

PLATINUM: satisfies the Gold requirements, has machine readable provenance documentation, uses unique identifiers in the data, the publisher has a communications team offering support. This is an exceptional example of an information infrastructure.

5-star deployment scheme for Open Data

Sources: Open data institute (UK), Centre for open science (US), Tim-Berners Lee
DANS: FAIR badge scheme

- First Badge System based on the FAIR principles: proxy for data quality assessment
- Operationalise the original principles to ensure no interactions among dimensions to ease scoring
- Consider Reusability as the resultant of the other three:
  - the average FAIRness as an indicator of data quality
  - \((F+A+I)/3 = R\)
- Manual and automatic scoring

2 User Reviews
1 Archivist Assessment
24 Downloads
First we attempted to operationalise R – Reusable as well... but we changed our mind

**Reusable** – is it a separate dimension? Partly subjective: it depends on what you want to use the data for!

<table>
<thead>
<tr>
<th>Idea for operationalization</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1. plurality of accurate and relevant attributes</td>
<td>≈ F2: “data are described with rich metadata” → F</td>
</tr>
<tr>
<td>R1.1. clear and accessible data usage license</td>
<td>→ A</td>
</tr>
<tr>
<td>R1.2. provenance (for replication and reuse)</td>
<td>→ F</td>
</tr>
<tr>
<td>R1.3. meet domain-relevant community standards</td>
<td>→ I</td>
</tr>
<tr>
<td>Data is in a TDR – unsustained data will not remain usable</td>
<td>Aspect of Repository → Data Seal of Approval</td>
</tr>
<tr>
<td>Explication on how data was or can be used is available</td>
<td>→ F</td>
</tr>
<tr>
<td>Data is automatically usable by machines</td>
<td>→ I</td>
</tr>
<tr>
<td>Findable (defined by metadata (PID included) and documentation)</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1. No PID nor metadata/documentation</td>
<td></td>
</tr>
<tr>
<td>2. PID without or with insufficient metadata</td>
<td></td>
</tr>
<tr>
<td>3. Sufficient/limited metadata without PID</td>
<td></td>
</tr>
<tr>
<td>4. PID with sufficient metadata</td>
<td></td>
</tr>
<tr>
<td>5. Extensive metadata and rich additional documentation available</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accessible (defined by presence of user license)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Metadata nor data are accessible</td>
</tr>
<tr>
<td>2. Metadata are accessible but data is not accessible (no clear terms of reuse in license)</td>
</tr>
<tr>
<td>3. User restrictions apply (i.e. privacy, commercial interests, embargo period)</td>
</tr>
<tr>
<td>4. Public access (after registration)</td>
</tr>
<tr>
<td>5. Open access unrestricted</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interoperable (defined by data format)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Proprietary (privately owned), non-open format</td>
</tr>
<tr>
<td>2. Proprietary format, accepted by Certified Trustworthy Data Repository</td>
</tr>
<tr>
<td>3. Non-proprietary, open format = ‘preferred format’</td>
</tr>
<tr>
<td>4. As well as in a preferred format, data is standardised using a standard vocabulary (for the research field to which the data pertain)</td>
</tr>
<tr>
<td>5. Data is additionally linked to other data to provide context</td>
</tr>
</tbody>
</table>
Creating a FAIR data assessment tool

Using an online questionnaire system
Website FAIRDAT

To contain FAIR data assessments from any repository or website, linking to the location of the data set via (persistent) identifier

The repository can show the resultant badge, linking back to the FAIRDAT website

Neutral, Independent
Analogous to DSA website

Mockups!
Display FAIR badges in any repository (Zenodo, Dataverse, Mendeley Data, figshare, B2SAFE, ...)

Mockups!
Can FAIR Data Assessment be automatic?

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Automatic? Y/N/Semi</th>
<th>Subjective? Y/N/Semi</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>S</td>
<td>S</td>
<td>Insufficient metadata is subjective</td>
</tr>
<tr>
<td>F3</td>
<td>S</td>
<td>S</td>
<td>Sufficient metadata is subjective</td>
</tr>
<tr>
<td>F4</td>
<td>S</td>
<td>S</td>
<td>Sufficient metadata is subjective</td>
</tr>
<tr>
<td>F5</td>
<td>S</td>
<td>S</td>
<td>Rich metadata is subjective</td>
</tr>
<tr>
<td>A1</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>I1</td>
<td>S</td>
<td>N</td>
<td>Depends on list of proprietary formats</td>
</tr>
<tr>
<td>I2</td>
<td>S</td>
<td>S</td>
<td>Depends on list of accepted formats</td>
</tr>
<tr>
<td>I3</td>
<td>S</td>
<td>S</td>
<td>Depends on list of archival formats</td>
</tr>
<tr>
<td>I4</td>
<td>N</td>
<td>S</td>
<td>Depends on domain vocabularies</td>
</tr>
<tr>
<td>I5</td>
<td>S</td>
<td>N</td>
<td>Depends on semantic methods used</td>
</tr>
</tbody>
</table>

Optional: qualitative assessment / data review
Focus: the I of FAIR

To be Interoperable (according to FORCE11 FAIR group):
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.

Note:
- Criteria apply to both data and meta data (data descriptions)
- I1 and I2 have to do with a common understanding (semantics) of the data
- I3 has to do with linking data
Interoperability is a very challenging term. My major concern here is that a data object is not interoperable per se or in an absolute manner. Interoperability is a property of two (or more) systems, it is achieved once the involved systems manage to exchange and use the target object. A detailed discussion on this occurred in the context of the DL.org European Project and a model aiming at discussing interoperability was documented by http://goo.gl/79xYV (here is the extended version http://goo.gl/fBU2sX).

In addition to that, interoperability is not a yes/no property. Rather there are several levels of "interoperability" that can be achieved in a given context.

Although "Interoperable" is a very appealing term I suggest to not use it. Rather I propose to use **Intelligible** (this will not change the acronym).
‘[...] a data object is not interoperable per se or in an absolute manner. Interoperability is a property of two (or more) systems, it is achieved once the involved systems manage to exchange and use the target object. [...] In addition to that, interoperability is not a yes/no property. Rather there are several levels of "interoperability" that can be achieved in a given context.

Although "Interoperable" is a very appealing term I suggest to not use it. Rather I propose to use **Intelligible** (this will not change the acronym).’

**Leonardo Candela** - National Research Council of Italy - Institute of Information Science and Technologies (CNR - ISTI)
Tim Berners-Lee 5 Star Deployment Scheme for Linked Open Data

★
Available on the web (whatever format) *but with an open licence, to be Open Data*

★★
Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★
As (2) plus non-proprietary format (e.g. CSV instead of Excel)

★★★★
All the above plus, use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★
All the above, plus: Link your data to other people’s data to provide context

Note: 1 star has to do with access license; 2-4 stars have to do with (standard) data formats; 5 stars has to do with links between data

https://www.w3.org/DesignIssues/LinkedData.html
### DANS ideas for measuring data interoperability

| ★ | Proprietary (privately owned), non-open format |
| ★★ | Proprietary format, but accepted by Certified Trustworthy Data Repository (‘accepted format’) |
| ★★★ | Non-proprietary, open format = ‘preferred format’ or ‘archival format’ |
| ★★★★ | As well as in a preferred format, data is standardised using a standard vocabulary (for the research field to which the data pertain) |
| ★★★★★ | Data is additionally linked to other data to provide context |

**Note:** we are struggling with two basic problems:
- How to score multi-file data sets, which have different levels of “Interoperability”
- Perhaps it is enough to distinguish between ‘proprietary, non-open’ and ‘preferred/archival format’
Further complications concerning data interoperability

- Domain vocabularies, thesauri, ontologies, classification schemes, etc. exist in a multitude of forms, having very different levels of applicability and acceptance
  - For instance, how to evaluate the degree of harmonization if in a database of 100 variables two use standard definitions

- How to incorporate the divergent levels of technological skills among users, which influence their preferences with respect to interoperability?
  - Example: users preferring a data set to be in Excel rather than in RDF, because they know how to use Microsoft and do not know how to use semantic web technologies
Questions for discussion

1. What are the indispensable aspects of interoperability of research data?
2. Is it indeed better to replace “interoperable” by “intelligible”, as Leonardo Candela suggested?
3. Interoperability seems to be rather a measure of degree than a matter of yes/no. If you agree, how to measure the degree of interoperability (or intelligibility)?
4. If data harmonization reflects a high level of (semantic) interoperability, how to establish the degree of harmonization in a data set?
5. How to take the different levels of user skills into account? How to acknowledge users that cannot work with Linked Open Data?
Thank you for listening!

“Tell us what you think!”

peter.doorn@dans.knaw.nl
www.dans.knaw.nl
http://www.dtls.nl/go-fair/