Applying the FAIR guiding principles to clinical data management and re-use

Stefan Schulz

Institute for Medical Informatics, Statistics and Documentation
Medical University
Auenbruggerplatz 2/V
8036 Graz (Austria)
https://purl.org/steschu
stefan.schulz@medunigraz.at

Berlin, 28 Nov 2017
FAIR principles

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

FAIR principles

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

- **F – Findable** – Enriching datasets with metadata and annotation to support high quality content retrieval
**FAIR principles**

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

- **F – Findable** – Enriching datasets with metadata and annotation to support high quality content retrieval
- **A – Accessible** – Facilitating access to the data according to clear regulation regarding licenses of use

FAIR principles

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

- **F – Findable** – Enriching datasets with metadata and annotation to support high quality content retrieval
- **A – Accessible** – Facilitating access to the data according to clear regulation regarding licenses of use
- **I – Interoperable** – Using machine-readable and internationally compatible standards for semantic annotations and metadata
- **R – Reusable** – Using exhaustive semantic annotations and metadata to reliably repurpose data, by preserving provenance, data production, and other contextual information.
FAIR principles

Manifesto for sustainable use of scientific research objects (data, workflows, algorithms) by humans and their digital agents

- **F** – **Findable** – Enriching datasets with metadata and annotation to support high quality content retrieval
- **A** – **Accessible** – Facilitating access to the data according to clear regulation regarding licenses of use
- **I** – **Interoperable** – Using machine-readable and internationally compatible standards for semantic annotations and metadata
- **R** – **Reusable** – Using exhaustive semantic annotations and metadata to reliably repurpose data, by preserving provenance, data production, and other contextual information.

*European Commission and the G20 encourage researchers to embrace the FAIR principles*

Current scope of FAIR: data-intensive science, optimised use of data acquired by public funding, idea of scientific data as a public good

What if we include also clinical "big" data?

- Images, lab (low / high-throughput), bio-signals
- All kinds of textual and coded data
- Patient-generated data

"FAIRify" data for primary and secondary use cases

- Use of clinical data for scientific research
- Other secondary uses (e.g. business intelligence)
- Improved primary use (e.g. decision support, personalised data visualisation, coding support)
Re-Defining the FAIR principles...

- ...for original patient data or data derived from patient data
- Contrasting current status with FAIR desiderata
- Requirements to implement FAIR principles for patient-related datasets:
  - Methods
  - Resources
  - Conditions
FAIR – Findability
FAIR – Findability

**Reality**
- Clinical data / documents identifiable and addressable even within closed systems difficult
- Retrieval of data from one patient or across several patients not supported by typical CIS
- Information retrieval across several CIS not supported
- No indexing of unstructured content

**Desiderata**
- Clinical data / documents are assigned a globally unique and eternally persistent identifier
- Both database and free text search facilitates quick content retrieval with a single CIS
- Meta-search across several CIS is supported
- Semantic indexing reduces impact of language variety
Patient? Pathologie?
rezent? rezidivierend?
Vesicoureteral reflux
Harnwegsinfekt? Hinterwandinfarkt?
<table>
<thead>
<tr>
<th>Colon-Ca</th>
<th>Kolon-Ca</th>
<th>Kolon-karzinom</th>
<th>Colon-carcinom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon-Karzinom</td>
<td>Kolonkrebs</td>
<td>Dickdarm-krebs</td>
<td>Dickdarm-Ca</td>
</tr>
<tr>
<td>Malignom des Kolon</td>
<td>Dickdarm-karzinom</td>
<td>Dickdarm-Ca</td>
<td>Bösartige Neubildung am Dickdarm</td>
</tr>
<tr>
<td>Bösartiger Dickdarm-tumor</td>
<td>maligne Neoplasie des Dickdarms</td>
<td>Karzinom des Dickdarms</td>
<td>maligne NPL des Colon</td>
</tr>
</tbody>
</table>
## Common misspellings

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Simvastatin</td>
<td>Sinvastatin</td>
<td>Simvastastin</td>
<td>Simvastain</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simvastad</td>
<td>Simbastatin</td>
<td>Simavstatin</td>
<td>Simavvastatin</td>
</tr>
<tr>
<td></td>
<td>Simastatin</td>
<td>Symvastatin</td>
<td>Simvastation</td>
<td>Simvaststin</td>
</tr>
<tr>
<td></td>
<td>Simvatatin</td>
<td>Simvatin</td>
<td>Simvatstain</td>
<td>Simvstatin</td>
</tr>
</tbody>
</table>
"Makroskopie: "Resektat nach Whipple": Ein noch nicht eröffnetes Resektat, bestehend aus einem distalen Magen mit einer kleinen Kurvaturlänge von 9,5 cm und einer großen Kurvaturlänge von 13,5 cm, sowie einem duodenalen Anteil von 14 cm Länge. 2 cm aboral des Pylorus zeigt die Dünndarmwandung eine sanduhrartige Stenose. Im Magen- und Duodenallumen reichlich zähflüssiger Schleim, sanguinolent; die Schleimhaut ist insgesamt livide. Auf lamellierenden Schnitten zähfestes weißliches, teilweise nodulär konfiguriertes Gewebe, ohne das Gallengänge manifest werden. Der distale Anteil des Ductus pankreaticus ist leicht erweitert und von der Papilla vateri aus 4,5 cm weit sondierbar, wobei er hier in einer peripankreatischen Narbenzone abbricht. Eine Gallengangsmündung läßt sich makroskopisch nicht abgrenzen. Die berichtete Duodenumstenose liegt 2,5 cm oral der Papilla vateri und steht mit der beschriebenen Narbenzone in direktem Zusammenhang.
FAIR – Accessibility
Accessibility

**Reality**
- Data are locked in silos, data import / export via costly custom procedures
- No transparent, secure, customisable authentication and authorization protocols
- Data access / exchange unclear. Bilateral agreements without a robust technical and regulatory framework. Informed consent for reuse of routine data missing
- Manual de-identification

**Desiderata**
- Data are accessible by their identifiers using a standardized, free, secure communication protocol.
- The protocol allows for authentication and authorization procedures
- Multidimensional access policies, dependent on de-identification, types and frequencies of data values, granularity, privacy regulations, informed consent
- Automated de-identification
De-identified narratives

Coded, de-identified extracts

- 82271004 | Injury of head region
- 125593007 | Facial injury
- 262749000 | Open wound of eyelid
- 313261004 | Open wound of chin
- 7771000 | Left side
- 255473004 | Symmetrical
- 51440002 | Right and left (qualifier value)
- 301939004 | Pupil constriction
- 255510006 | Slight
- 366084008 | Finding of ocular divergence
- 399054005 | Exotropia (disorder)
- 8966001 | Left eye
- 282977007 | Does bend
- 66019005 | Extremity
- 22253000 | Pain observations
- 122545008 | Stimulation
- 80447000 | Aqueduct of Sylvius
- 118592000 | Velocity
- 255473004 | Symmetrical
- 17621005 | Normal
- 168733007 | Standard chest X-ray normal
- 2004005 | Normotensive
FAIR – Interoperability / Reusability
<table>
<thead>
<tr>
<th>Reality</th>
<th>Desiderata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most content is in compact clinical language. Most structured content does not make use of semantic standards</td>
<td>Content is represented using internationally sharable and &quot;FAIRified&quot; computable formalisms and vocabularies</td>
</tr>
<tr>
<td>What clinical data means and how data are related limited to experts reading clinical text. Contexts are hidden, correct interpretation limited to insiders</td>
<td>Context (time, certainty, authorship, purpose) is made explicit</td>
</tr>
<tr>
<td>Data provenance often unclear</td>
<td>Data provenance allows estimation of data quality</td>
</tr>
<tr>
<td>Undefined licence regulations prevent data re-use</td>
<td>Data are released with clear usage licenses</td>
</tr>
<tr>
<td>Data integration hubs consolidate and integrate heterogeneous data</td>
<td>Data integration hubs consolidate and integrate heterogeneous data</td>
</tr>
</tbody>
</table>
Interoperability through common standards

<table>
<thead>
<tr>
<th>Code (SNOMED CT, LOINC)</th>
<th>Value</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>254730000</td>
<td>Superficial spreading malignant melanoma of skin</td>
<td>392521001</td>
</tr>
<tr>
<td>301889008</td>
<td>Excision of malignant skin tumor</td>
<td>392521001</td>
</tr>
<tr>
<td>47224004</td>
<td>Skin of posterior surface of lower leg</td>
<td></td>
</tr>
<tr>
<td>7771000</td>
<td>Left</td>
<td></td>
</tr>
<tr>
<td>258673006</td>
<td>Diameter</td>
<td>2.41</td>
</tr>
<tr>
<td>258403002</td>
<td>Lymph node level IV</td>
<td></td>
</tr>
<tr>
<td>94339008</td>
<td>Secondary malignant neoplasm of inguinal lymph nodes</td>
<td>15240007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2667000</td>
</tr>
</tbody>
</table>
Interoperability through computable semantics

Diagnose Organversagen

Organ
Herz

Status
Verdacht

Ursache
Ischämische Herzenerkrankung

Ja

... es besteht Verdacht auf Herzinsuffizienz verursacht durch die ischämische Herzerkrankung

Ja

Nein

k.A.

DIAGNOSIS

Suspected heart failure caused by ischaemic heart disease

Data provenance / context examples

- "94 kg"
  - reported by patient / measured in hospital
  - measured at admission / at discharge
  - target weight (e.g. after obesity treatment)
  - extracted from database / text mined from letter

- "G03.9 – Meningitis unspecified"
  - ICD code for billing – used for non-confirmed cases*
  - ICD code from death certificate
  - ICD code resulting from text mining

Challenges for clinical data "FAIRification"
Challenges for clinical data "FAIRification"

Source: https://de.toonpool.com/cartoons/Datenfriedhof_224540#img9
Standardisierung: Interoperabilität gewinnt an Fahrt

Dtsch Arztebl 2017; 114(47): A-2200 / B-1847 / C-1801

THEMEN DER ZEIT
Krüger-Brand, Heike E.

Die Einsicht wächst, dass grundlegende Standards für die Datenkommunikation im Gesundheitswesen unerlässlich sind. Einige Fortschritte gibt es zu vermelden.
Challenges for clinical data "FAIRification"

- Standards
  - Terminologies (SNOMED CT, LOINC, WHO classifications, ISO IDMP standards, ...)
  - Information models (HL7, FIHR)
  - Clinical documents (HL7-CDA, IHE XDS.b)

Medikationsplan Plus - http://egesundheit.nrw.de/projekt/medikationsplan-plus/
LOINC User Group Deutschland - http://www.loinc.de
Identification of medicinal products
Challenges for clinical data "FAIRification"

- Infrastructures and regulations
  - Data integration centres and digital health research platforms
  - Data safety and privacy policies
  - Involvement of all stakeholders

Consolidation of data sources allows for new research and treatment approaches in medicine
http://www.uni-mainz.de/presse/aktuell/2180_ENG_HTML.php
TMF – Arbeitsgruppe Datenschutz http://www.tmf-ev.de/Arbeitsgruppen_Foren/AGDS.aspx
Challenges for clinical data "FAIRification"

- **Resources**
  - Local interface dictionaries for structured data entry and text mining, linked to terminology standards (e.g. crowdsourcing approaches)
  - Large training data for supervised / unsupervised learning → improving performance of clinical information extraction
    - de-identified clinical documents
    - document fragments
    - n-gram statistics


S. Schulz - Building an experimental German user interface terminology linked to SNOMED CT SNOMED CT EXPO https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=45525419
Challenges for clinical data "FAIRification"

- **Collaborative users**
  - Improved, personalised user interfaces of electronic health records:
    - Better data quality, more efficient use, higher degree of structured and coded content

- **Basic research**
  - Research on clinical language, lexicology, knowledge acquisition from big data
  - Ontologies and knowledge representation

- **Translational research commons – bridging between health care and molecular biology**

Thank You!

Contact:

Stefan Schulz
purl.org/steschu
stefan.schulz@medunigraz.at
Ventrikelfunktionsstörung I.
rhythmusspezifischen
konnte am ##.##.### bei
verabreichten
RI
Kontrolle INR
elektrophysiologische Untersuchung.
Bei neuerlich auftretenden
#,, cm# und AINS I-II°
sind für
###/###ms
Ospexin #x#g
als weitgehend
und Zustand nach Implantation
mit erhöhtem Risiko op-tauglich
eines angiologischen
als Kind
HCT ##mg/###mg/##mg
kardiolog. Kontrolluntersuchungen beim FA für
Velpurin
Über der Lunge
Li. Vorhof einschließlich Herzohr frei
### Examples

<table>
<thead>
<tr>
<th>Clinical text</th>
<th>SNOMED CT concepts (FSNs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;... the duodenum. The mucosa is...&quot;</td>
<td>'Duodenal structure (body structure)'</td>
</tr>
<tr>
<td></td>
<td>'Mucous membrane structure (body structure)'</td>
</tr>
<tr>
<td></td>
<td>'Duodenal mucous membrane structure (body structure)'</td>
</tr>
<tr>
<td>&quot;... Hemorrhagic shock after RTA...&quot;</td>
<td>'Traffic accident on public road (event)'</td>
</tr>
<tr>
<td></td>
<td>'Traffic accident on public road (event)', 'Renal tubular acidosis (disorder)'</td>
</tr>
<tr>
<td></td>
<td>'Traffic accident on public road (event)' or 'Renal tubular acidosis (disorder)'</td>
</tr>
<tr>
<td>&quot;... travel history of suspected dengue...&quot;</td>
<td>'Suspected dengue (situation)'</td>
</tr>
<tr>
<td></td>
<td>'Suspected (qualifier value)'</td>
</tr>
<tr>
<td></td>
<td>'Dengue (disorder)'</td>
</tr>
</tbody>
</table>
Problem with large terminologies

<table>
<thead>
<tr>
<th>Metric</th>
<th>SNOMED CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept coverage [95% CI]</td>
<td></td>
</tr>
<tr>
<td>Text annotations – English</td>
<td>0.86 [0.82-0.88]</td>
</tr>
<tr>
<td>Term coverage [95% CI]</td>
<td></td>
</tr>
<tr>
<td>Text annotations – English</td>
<td>0.68 [0.64; 0.70]</td>
</tr>
<tr>
<td>Inter annotator agreement</td>
<td></td>
</tr>
<tr>
<td>Krippendorff's Alpha [95% CI]</td>
<td></td>
</tr>
<tr>
<td>Text annotations</td>
<td>0.37 [0.33-0.41]</td>
</tr>
</tbody>
</table>

(similar results with alternative annotation task, using non-SNOMED UMLS extract)

Ecosystem of semantic assets

Information Models

Process Models

Terminologies

Guideline Models
...describe and standardize a neutral, language-independent sense
- The meaning of domain terms
- The properties of the objects that these terms denote
- Representational units are commonly called “concepts”
- RTs enhanced by formal descriptions = "Ontologies"
Information Models

Guideline Models

Core Reference Terminology

Aggregation Terminologies (Classifications)

- Systems of non-overlapping classes in single hierarchies, for data aggregation and ordering.
- aka classifications, e.g. the WHO classifications
- Typically used for health statistics and reimbursement
• Reference and aggregation terminologies represent / organize the domain
• They are not primarily representations of language
• They use human language labels as a means to univocally describe the entities they denote, independently of the language actually used in human communication

• Systems of non-overlapping classes in single hierarchies, for data aggregation and ordering.
• aka classifications, e.g. the WHO classifications
• Typically used for health statistics and reimbursement
User Interface Terminology (language specific)

- Collections of terms used in written and oral communication within a group of users
- Terms often ambiguous.
- Entries in user interface terminologies to be further specified by language, dialect, time, sub(domain), user group.
User Interface Terminology (e.g. Portuguese)

"Ca"
"Cálcio"

[chemistry]

"Ca"
"Câncer"
"Carcinoma"

[oncology]

Reference Terminology

5540006 | Calcium (substance) |

68453008 | Carcinoma (morphologic abnormality) |
MUG-GIT: Creation of German Interface Terminologie for SNOMED CT

Raw full terms (DE)

Human Validation
- dependent on use cases
- e.g. input for official translation
- e.g. starting point for crowdsourcing process for interface term generation
- lexicon for NLP approaches

All SCT descriptions (EN)

Non-Translatable SCT descriptions

Filter concepts with identical terms across translations

Phrase generation rules

Term reassembling heuristics

Human curation
- correct most frequent mis-translations
- remove wrong translations
- check POS tags
- normalise adjectives
- add synonyms

Curated ngram translations (DE)

Clinical corpus (DE)

n-grams (DE)

New Token translations

Rule exec

Char translation rule acquisition

Reference corpus (DE)

Token translations

untranslated tokens

POS tags

Non-Translatable SCT descriptions

Chunker

Rules

Translatable SCT descriptions (EN)

n-grams (EN)

Raw full terms (DE)

Google Translate

Token translations

untranslated tokens

POS tags

n-grams (EN)

n-grams (EN)

Translatable SCT descriptions (EN)
### ngram – core vocabulary

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
<th>German Translation</th>
</tr>
</thead>
</table>
| vaginal                                   | 1         | vaginales | Scheiden-
| fluoroscopic guidance                     | 2         | Durchleuchtungskontrolle | FF |
| disc                                      | 1         | Scheibe | FF |
| lower limb                               | 2         | unteres | Bein |
| brain                                     | 1         | Gehirn | Hirn | Encephalon |
| preparation                              | 1         | Zubereitung | Aufbereitung | Präparation |
| method                                    | 1         | Verfahren | Methode |
| of bone                                   | 2         | des Knochens | _Knochen_ |
| Red                                       | 1         | rotes | JJ |
| Monitoring                                | 1         | Überwachung | Monitoring |
| Computed                                  | 1         | berechnetes | Computer |
| phalanx                                   | 1         | Phalanx | NN | F |
| subsp.                                    | 1         | Phalanx | NN | F |
| anastomosis                               | 1         | Anastomose | Anastomosierung |
| vessel                                    | 1         | Blutgefäß | Gefäß |
| Computed tomography                       | 2         | Computertomographie | |
| uterus                                    | 1         | Uterus | Gebärmutter |
| difficulty                                | 1         | Schwierigkeit | |
| elbow                                     | 1         | Ellbogen | Ellbogengelenk |
| high                                      | 1         | hohes | JJ |
| food                                      | 1         | Lebensmittel | Speise | Nahrungsmittel |
| Observation                               | 1         | Beobachtung | |
| using fluoroscopic                        | 2         | unfähig zu |
| unable                                    | 1         | unfähiges | JJ |
| Peripheral                                | 1         | peripheres | JJ |
| unable to                                 | 2         | unfähig zu |
| Vascular                                  | 1         | vaskuläres | JJ |
| using fluoroscopic guidance               | 3         | mit Durchleuchtungskontrolle |
| Benign neoplasm                           | 2         | gutartiges | benignes | |

The table provides information on the frequency and German translations of various terms related to medical topics such as anatomy, diagnostic procedures, and medical conditions.
<table>
<thead>
<tr>
<th>Date and Code</th>
<th>Code</th>
<th>Description</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>20170315_240011_002</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneubildung</td>
</tr>
<tr>
<td>20170315_240011_003</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neubildung des Hirns</td>
</tr>
<tr>
<td>20170315_240011_004</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneubildung</td>
</tr>
<tr>
<td>20170315_240011_005</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Gehirns</td>
</tr>
<tr>
<td>20170315_240011_006</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneoplasie</td>
</tr>
<tr>
<td>20170315_240011_007</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Hirns</td>
</tr>
<tr>
<td>20170315_240011_008</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneoplasie</td>
</tr>
<tr>
<td>20170315_240011_009</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Gehirns</td>
</tr>
<tr>
<td>20170315_240011_010</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneoplasie</td>
</tr>
<tr>
<td>20170315_240011_011</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Hirns</td>
</tr>
<tr>
<td>20170315_240011_012</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneoplasie</td>
</tr>
<tr>
<td>20170315_241010_001</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neubildung des Großhirns</td>
</tr>
<tr>
<td>20170315_241010_002</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neoplasie des Großhirns</td>
</tr>
<tr>
<td>20170315_241010_003</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neoplasie des Großhirns</td>
</tr>
<tr>
<td>20170315_242015_001</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neubildung des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_002</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neubildung des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_242015_003</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_004</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_242015_005</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_006</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_243013_001</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neubildung des Temporalappens</td>
</tr>
<tr>
<td>20170315_243013_002</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neubildung des Lobus temporalis</td>
</tr>
<tr>
<td>20170315_243013_003</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasie des Temporalappens</td>
</tr>
<tr>
<td>20170315_243013_004</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasie des Lobus temporalis</td>
</tr>
<tr>
<td>20170315_243013_005</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasie des Temporalappens</td>
</tr>
</tbody>
</table>