**Answer to Question: 1**

Disorders:

Pancreatitis - 75694006

Pancreatic Cancer - 363418001

Symptoms:

Abdominal Pain - 21522001

Loss of Appetite - 79890006

Body Structures

Gall Bladder - 28231008

Both upper arms - 6973007

Pharma/Bio Products

Omeprazole - 25673006

Capecitabine - 108761006

Procedures

Amylase Measurement - 64435009

Ultrasound - 16310003

**Answer to Question: 2**

It appears that both web services have gone to great lengths to maintain all the FAIR principles in their work. It’s important to note that both services ultimately implement the system *SNOMED CT*, as such the faults and benefits of the system are reflected in both works.

It would help perhaps to - briefly - review the FAIR principles and to examine where either service manages or fails to uphold these standards. I’ve opted to use the guiding principles of FAIR data described in the force11 website as it comes directly recommended from the Data FAIRport main page.

By *Findable*, it is entailed that data is described with rich metadata. Data and metadata are then assigned with a globally unique and persistent identifier. Data/metadata is then registered or indexed in a searchable resource, and metadata specifies the data identifier.

Both systems follow these guiding principles. The (meta)data is rich, and easily findable though each respective website, either through their search engine or by ID.

*Accessible* data refers to (meta)data that is retrievable by ID using a standardized communications protocol. As previously discussed this is most naturally true for both services.

The protocol is open, free and universally implementable. Also true, evidenced by the fact that the source code for both services is presented below. The following points however cannot be tested, this does not mean that both systems have failed to uphold them: “The protocol allows for an authentication and authorization procedure, where necessary” and “metadata are accessible, even when the data are no longer available”.

*Interoperability* would perhaps be the one point where both services fail, by interoperable it is expected that (meta)data is organized in a formal, accessible, shared and broadly applicable language for knowledge representation. The vocabulary of said language must in itself follow FAIR principles and finally (meta)data must include qualified references to other (meta)data. Even if we assume English to be the de facto Lingua Franca of the medicine, we can’t very well say that the vocabularies used to describe ontology follow FAIR principles, since they pre-date them.

Finally we come to *Re-Usability*, the Force11 describes Re-Usability as:

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.

All of these points are demonstrably true.

Sources (Other than the ones provided by the question):

<http://www.datafairport.org/>

<https://www.force11.org/group/fairgroup/fairprinciples>

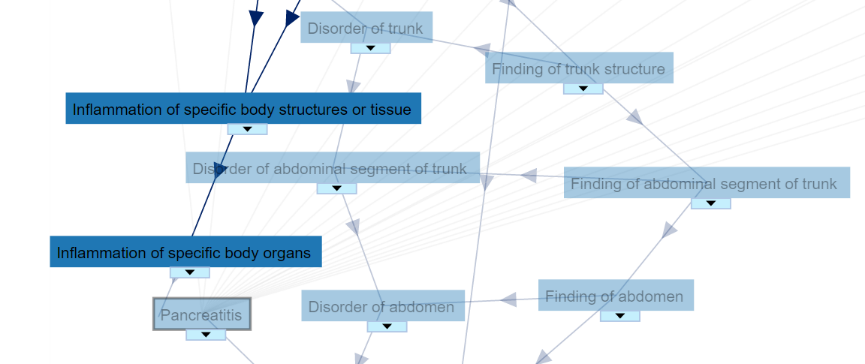
<https://github.com/ncbo>

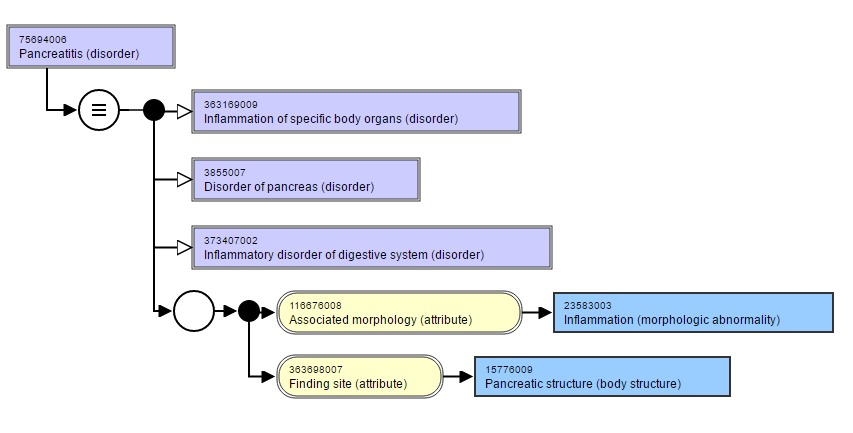
<https://github.com/IHTSDO/sct-browser-frontend>

**Answer to Question: 3**

*BioPortal* is a web service wherein biomedical ontological terms are stored in a free and readily available database. Their own mission probably states it best, although with a hint of exaggeration: “*the world’s most comprehensive repository of biomedical ontologies”*.

In both *BioPortal* and the *IHSTDO* implementation split their various terms in a hierarchy of parents and modules, not dissimilar to classes in *Object Oriented Programming*. For examples sake, picture *1* and *2* illustrate the difference in between both implementations.





2 - IHSTDO

1 - BioPortal

Cyst of pancreas (disorder) [31258000]

Parents:

|  |  |
| --- | --- |
| SNOWMED CT | ICD-10-CM |
| Cyst of pancreas (disorder) [31258000] | Cyst of pancreas [K86.2] |
| Disorder of pancreas (disorder) [3855007] | Disease of pancreas, unspecified [K86.9] |
| Cyst of abdomen (disorder) [446064000] | Cyst of abdomen , Intra-abdominal and pelvic swelling, mass and lump, unspecified site [R19.00] |
| Mass of pancreas (finding) [425810000] | Mass of pancreas , Other intra-abdominal and pelvic swelling, mass and lump [R19.09] |

**Answer to Question: 4**

1. After some minor installation problems – specifically *ant* wasn’t cooperating with the perfectly adequate JAVA\_HOME path – there were no major setbacks when it comes to executing the program. One should keep in mind however that this program was clearly designed for users with advanced computer literacy, there’s no such thing as a graphics interface, one quickly finds oneself investigating which types of formatting the input file finds adequate and the previously mentioned JAVA\_HOME path error certainly took far longer to solve than it ever should.

Nevertheless, there are some issues with the program that are inexcusable, even assuming that the end user is a computer engineer. After trying various format styles I’ve found, probably erroneously, that they serve no purpose other than separating paragraphs and avoiding compilation errors relating to carriage returns.

1. After executing the command I was able to produce the text files annexed bellow:

./ApplyDNorm.sh config/banner\_NCBIDisease\_UMLS2013AA\_TEST.xml data/CTD\_diseases.tsv output/simmatrix\_NCBIDisease\_e4.bin AB3P\_DIR TEMP angela-fixed.txt text.txt

(angela-fixed.txt and text.txt represent input and output respectfully)

1. As with all problems relating to natural language the crux of the matter lies with context.

After downloading the gigabytes of data necessary for the execution of *Ab3P*, I found that the resulting output remained much the same. In fact it probably got worse, as the system identified the English National Health Service abbreviature – NSH – as some sort of disease.

Natural Language Understanding, and its related subfields, are ultimately an AI-hard problem, quite frankly it’s a problem that is far too complex to be solved by an improvement of DNorm’s strategy or teaching algorithms.

1. I have to admit I couldn’t for the life of me get the TaggerOne code to work. As such I had to content myself with the demonstration site found on the download link. I have annexed some pictures of the result bellow. What poor findings I managed to collect were even less encouraging, as the narrative text would often times work with certain sections of text, but not its entirety and the PMID would not work at all.

Additional Sources:

Formatting guide used for DNorm and TaggerOne:

<https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/Format.html>

AI-Complete (or AI-Hard) reference (caution: Wikipedia)

<https://en.wikipedia.org/wiki/AI-complete>

**Annex:**

**angela-fixed.txt :**

20085714|t|This is a title

20085714|a|My GP had referred me because several blood tests showed that my amylase levels had remained significantly raised months after my gall bladder was removed. The consultant said there was nothing wrong with me, my amylase level was only mildly raised and there are lots of people walking […] do suffer, more people could be diagnosed earlier, with better outcomes.

The Full text was cut as it is quite large.