

```
1.21
notwriter
38.p12
ession NCBI Assembly:GCF_000001405.38
26 March 2018
a NCBI Homo sapiens Annotation Release 109
VC_000001.11 1 248956422
www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.d
fSeq region 1 248956422
h:9606;Name=1;chromosome=1;gbkey=Src;genome=cl
stRefSeq pseudogene 11874 14409
heID:100287102,HGNC:HGNC:37102;Name=DDX11L1;d
ke
DDX11L1;gene_biotype=transcribed_pseudogene;p
stRefSeq transcript 11874 14409
e0;Dbxref=GeneID:100287102,Genbank:NR_046018.
ike 1-transcript_id=NR_046018.2
```

AUG 08, 2023

OPEN  ACCESS



#### DOI:

[dx.doi.org/10.17504/protocols.io.q26g7p4r1gwz/v1](https://dx.doi.org/10.17504/protocols.io.q26g7p4r1gwz/v1)

#### Protocol Citation:

Yo Yehudi, Caroline Jay, lukasnoehrer, Carole Goble 2023. Subjective Data Models in Bioinformatics - Interview-based personal data model elicitation. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.q26g7p4r1gwz/v1>

#### MANUSCRIPT CITATION:

<https://doi.org/10.48550/arXiv.2208.12346>

**License:** This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

## Subjective Data Models in Bioinformatics - Interview-based personal data model elicitation

Carole

Yo Yehudi<sup>1,2</sup>, Caroline Jay<sup>2</sup>, lukasnoehrer<sup>2</sup>, Goble<sup>2</sup>

<sup>1</sup>OLS; <sup>2</sup>University of Manchester

Subjective Data Models



Yo Yehudi

OLS

**Protocol status:** Working  
We use this protocol and it's working

**Created:** Jul 20, 2023

**Last Modified:** Aug 08, 2023

**PROTOCOL integer ID:**  
85279

**Keywords:** bioinformatics,  
data models, subjective data  
models

## ABSTRACT

Biological science produces large amounts of data in a variety of formats, which necessitates the use of computational tools to process, integrate, analyse, and glean insights from the data. Researchers who use computational biology tools range from those who use computers primarily for communication and data lookup, to those who write complex software programs in order to analyse data or make it easier for others to do so. This research examines how people differ in how they conceptualise the same data, for which we coin the term "subjective data models".

This protocol provides detailed steps to elicit an individual's perceptions of biological data models. This is accomplished by interviewing individuals with interdisciplinary backgrounds: academic-level biological experience and varied levels of computational experience. Participants in the study are given three tasks, alongside a set of props to form the experiment.

Results expected: This set of tasks is particularly useful for PIs, software engineers, bioinformaticians, and other researchers in biological/computational science who design data models or graphical/command line interfaces. It provides the researchers with a real-world idea of how their users and potential users conceptualise data models for biological data, as well as providing information about the FAIRness, usability and context of identifiers used in the individual biological file formats.

### Tasks:

1. A set of cards with biological data model entities and properties, and asked to "think aloud" whilst organising the cards in a way that makes sense to them.
2. A set of biological files (GFF and FASTA), where participants are asked to map terms on the cards to terms on the files.
3. Defining an "entry point" into the model.

### Tools provided:

Preparation: an interview guide, a background survey for each participant to complete, covering the computational tools they have used in the past, formal education and degree level, file formats and programming languages a participant is familiar with.

Interview props: a list of suggested data model cards for participants to sort, files to use as an example of mapping a data model to real-world files.

## IMAGE ATTRIBUTION

CC-BY 4.0 Yo Yehudi

## MATERIALS

Web cam or microphone, laptop, or other recording device.

Cards (for writing or printing the data model entities).

Pen, paper.

## BEFORE START INSTRUCTIONS

Before initiating this study, ensure you have recording equipment, pen, paper, small cards (business card or post-it size recommended), and a large table space in a sufficiently quiet / private room for 20-40 minutes per session.

### Initial experiment set-up

- 1 Start by preparing the materials you'll need for your interviews.

10m

Gather the following items:

- 1.1 **Small cards**, no larger than a business card. Index cards are generally too big - we tried it! We recommend having at least two colours of cards, in contrasting colour-blind friendly tones (that means NOT red and green, as red/green colourblind is the most common type of colourblindness, found in around 10% of men).

A couple of good choices:

- Pink and Yellow
- Yellow and Blue

If you aren't colourblind yourself, ask a friend or use an online colourblind checking tool to make sure the card colours are accessible.

- 1.2 **Recording equipment**, such as a laptop with a decent-quality camera, or a smartphone and a recording stand.

- 2 **Print biological files for the file-mapping task**

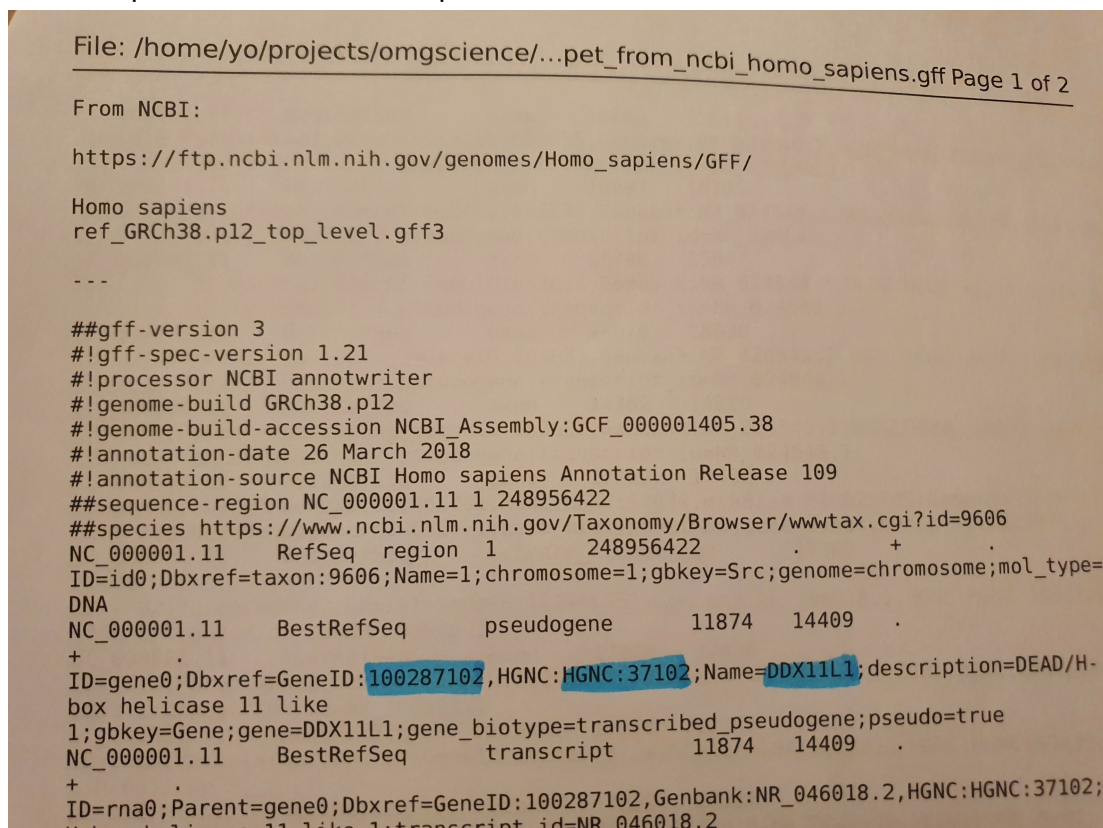
These could be the pre-supplied files (GFF and FASTA for Homo sapiens and Drosophila melanogaster), or other files of your choosing that apply to your domain.

Note that you will have to **change the terms that go on the cards** if you choose your own file - see

step 3.1 for more info.

## 2.1 Highlight terms from the files

We highlighted a selection of terms in the files that we believed mapped to the data model cards we provided to users. Example screenshot:



```
File: /home/yo/projects/omgscience/...pet_from_ncbi_homo_sapiens.gff Page 1 of 2
From NCBI:
https://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/GFF/
Homo sapiens
ref_GRCh38.p12_top_level.gff3
---
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build GRCh38.p12
#!genome-build-accession NCBI_Assembly:GCF_000001405.38
#!annotation-date 26 March 2018
#!annotation-source NCBI Homo sapiens Annotation Release 109
##sequence-region NC_000001.11 1 248956422
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606
NC_000001.11 RefSeq region 1 248956422 . +
ID=id0;Dbxref=taxon:9606;Name=1;chromosome=1;gbkey=Src;genome=chromosome;mol_type=DNA
NC_000001.11 BestRefSeq pseudogene 11874 14409 .
+
ID=gene0;Dbxref=GeneID:100287102,HGNC:HGNC:37102;Name=DDX11L1;description=DEAD/H-box helicase 11 like
1;gbkey=Gene;gene=DDX11L1;gene_biotype=transcribed_pseudogene;pseudo=true
NC_000001.11 BestRefSeq transcript 11874 14409 .
+
ID=rna0;Parent=gene0;Dbxref=GeneID:100287102,Genbank:NR_046018.2,HGNC:HGNC:37102;
H-box helicase 11 like 1;transcript id=NR_046018.2
```

Above: Preview of homo sapiens gff file, with three terms highlighted in bright blue.

## 2.2 If you're using the pre-provided files

Highlight the following terms:

A	B
File	Term(s) to highlight
homo_sapiens.gff	100287102
	HGNC:37102
	DDX11L1
flybase_d_melanogaster.gaf	FBgn0043467
	GO:004819

A	B
ncbi_homo_sapiens.gene_info	9606
	1
	A1BG
	MIM:138670
	HGNC:HGNC:5
	Ensembl:ENSG 00000121410

Table of highlighted terms in each file

(Note that the files provided are plain-text files, which is why the highlighting is required after it has been printed)

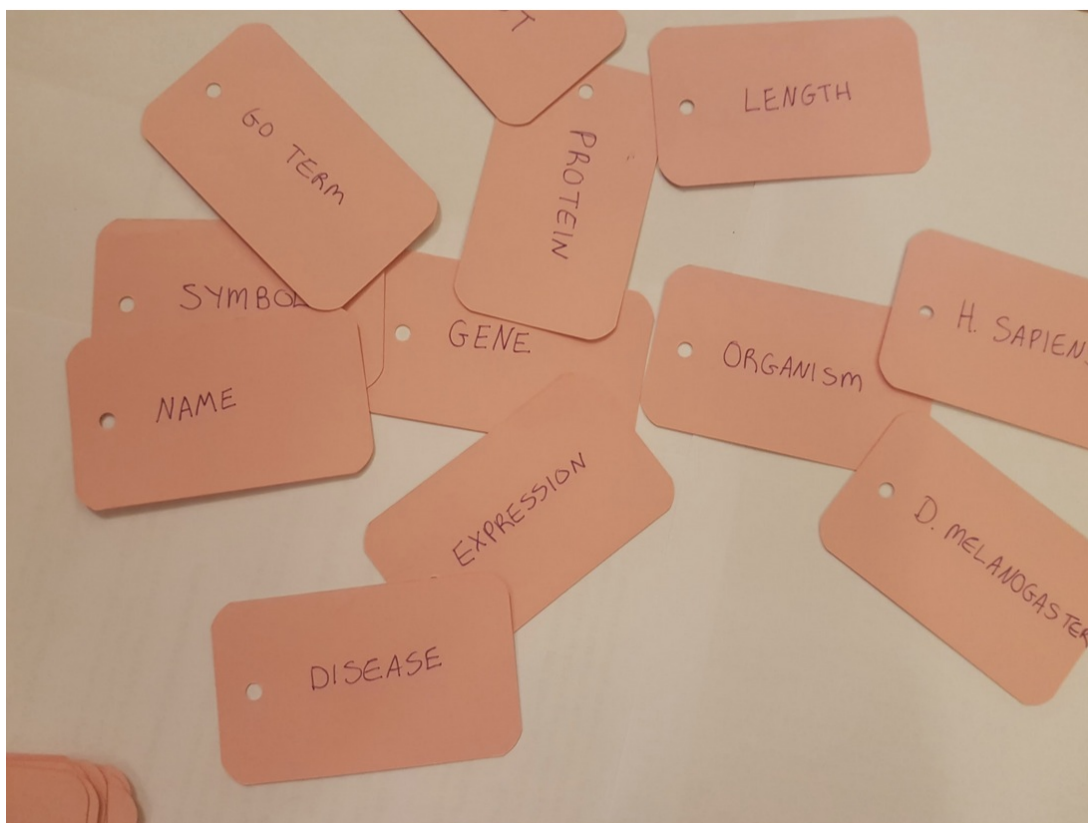
### 2.3 If you're using your own files

*(If you're using the pre-provided files, skip this step).*

You'll need to select a few terms in your files to highlight in some way for the card mapping exercise. Make sure that they're terms that *you* would link to the cards in your data model. Also see step 3.1 for custom cards associated with these files.

## 3 Print or hand-write terms on the cards

Print or write exactly one data model term from the cards file per card. It should look something like this:



Above: Image of several cards showing one term from cards.txt on each card.

If your handwriting is poor, we recommend using a printer. You could print out stickers and stick them onto cards, or print on thicker paper/thin card, and cut the cards out.

Alternatively, try finding a friend who can write neatly!

### 3.1 If you're using your own files:

*(If you're using the pre-provided files, skip this step).*

If you use your own files, we suggest you remove some or all of the following cards from cards.txt, and instead choose your own terms to write on the cards:

A
ASTHMA
BRCA1
BRCA1_HUMAN
D. MELANOGASTER
DIABETES
DISEASE
GO:0005515 PROTEIN BINDING

A
H. SAPIENS
P53
PMID:13791489
Q9H4C3_HUMAN
XY

Above: Table of terms that are specific to the pre-supplied files.

Consider providing alternative but equivalent terms instead.

For example:

- If you are using a **genomic/proteomic dataset about zebrafish**, you might replace "BRCA1" with a zebrafish gene symbol instead, and BRCA1\_HUMAN with a zebrafish protein accession.
- If your **dataset doesn't touch on diseases**, you might choose to replace the terms "disease", "asthma", and "diabetes" with something more relevant to your domain.

## "Dry run" time

- 4 Once you have all your files set up, you'll want to try the study once or twice before doing it "officially". You might learn some things about the way you set the table up, the colours, words, or sizes of your cards, the files you chose, or something else. Find someone who would be eligible to participate in your study who is willing to give it a go, and give you feedback if they have any questions, confusion, or doubts.

*Examples of changes from a pilot interview:*

- *When we piloted this study the first time, we used HUGE index cards, and often ran out of space on the table. Changing to cards the size of business cards helped a lot!*
- *We also chose to use a few terms on cards that we changed, because they made sense to us... but not to anyone else.*
- *Another change we could have implemented (if it wasn't a bit of a logistical challenge) would have been to bring a huge sheet of paper (think A0 poster size) and markers. Some participants REALLY wanted to draw network/graph like diagrams. We didn't do it in the end because travelling with poster-sized paper pads on the train (to interview venues) was not very feasible.*

Depending on how it goes (do you have to make a lot of changes?) and whether or not you need a full ethics review, you might choose not to count the dry runs as part of the primary dataset.

Check out the next section for tips on how to run the interviews, even the "dry runs".



## Batching your interviews

- 5** If you can, consider doing the interviews in batches. Go to a conference, or find a friend or colleague at a local friendly institute, book a meeting room, and set up a few slots throughout the day (no shorter than one hour!). This saves effort on set-up time.

Make sure you have quite a few copies of your information sheet and consent form printed out for participants. Send these to participants in advance, but be prepared to administer them on the day if they didn't read them before the appointment.

## Running the "real" interviews

**6 Set up the stage:**

Lay your cards out on the table - in a pile, not in any particular order. Make sure your recording device is on, and if relevant, pointed at the table. Check the audio can be heard, if the room is noisy.

Put down spare sheets of paper and pens, for the participants to use if desired.

**7 Consent and information sheets**

Before going any further, make sure the participant understands what the purpose of the study is, roughly what will happen, and if there are any consent or information forms to share with them - now is the time to get participants to read and sign them, so they know what you'll be doing with their data and recordings.

**8 Start recording**

Once the paperwork is done, it's time to kick off! Start the recording, double-check it's covering the correct area and that you'll be able to read / understand the cards and/or audio afterwards.

Make a note of which task(s) you present to the user first - tasks A, B, and C could be in any order and should vary between participants roughly equally.

**8.1 Task A - card sorting**

To participant: "Okay, we're going to run through a card sorting task. I have a set of cards here with some biological data-related words on them. Can I get you to sort or arrange them in a way that makes sense to you? There aren't any right or wrong answers here – I want to know what you think.

I also have some pens and paper – if you want to sketch, make notes, or create your own cards, feel free."

Actions:

1. Provide participant with the cards, blank cards, paper, and pens.



## 2. Observe the participant for cues

Possible (contextual) prompts / discussion points during the card sort:

- “Are you looking for anything in particular?”
- “You can add your own cards if it helps”
- (Explain what a certain card means, if the user is unfamiliar with it or asks questions – e.g. “What is a GO Term?”)
- “You can put any cards you don’t understand to the side – that’s useful for me to know, too”
- “Do you think there are any relationships between the cards you’ve sorted?”
- “Why did you choose to group these together?”

Once the card sort appears to be nearing a final stage:

“Are there any cards (or piles/groups of cards) you think are more important or interesting than the others?”

## 8.2 Task A Tidy-up

“Okay, thanks – this was really interesting. I’m just going to take a few quick photos of this, if you don’t mind?”

If there were any hand-drawn cards or sketches: “Is it alright to photograph these too? Thanks!”

1. Photograph all outputs of the card sort, assuming the participant agrees.
  - a. If any cards are arranged in piles, make sure to photograph cards that are underneath the top layer.
  - b. If the participant has hand-written anything, make sure that their handwriting is clear enough to be read, or clarify anything that isn’t readable.
  - c. Ensure that the photographs are of sufficient quality / focus to be read later on.
2. Remove any hand-written cards or sketches. Ideally shred as soon as possible.

## 9 Task B - file mapping

“Okay, nearly done now! One final thing left to do. I have some copies of biological files here, with a few sections highlighted. Can I get you to match the highlighted sections to items on the cards? Just write the card name (if any) on the paper, or tell me which cards it matches to.”

Actions:

1. Provide the participant with paper copies of your chosen files with highlighted terms
2. (Afterwards, if relevant) Photograph all outputs, assuming the participant agrees.

a. If the participant has hand-written anything, make sure that their handwriting is clear enough to be read, or clarify anything that isn't readable.

## 10 Task C - entry points

*This task should be done **after** Task A or Task B, once the participant has some familiarity with the cards.*

"Can you please identify an "entry point" into the data model, in the cards? This is somewhere that you might start if you had a research task or experiment to perform."

If relevant, ask the participant to explain why they chose the card(s).

## Wrap-up

## 11

Review all your work, check if the participant has any questions, and let the participant know what your future plans are - e.g. will you be writing up the study, will they be informed about the outcome?