Integrating and querying data with AskOmics

#### Anthony Bretaudeau, Olivier Dameron, Olivier Filangi, Xavier Garnier, Fabrice Legeai

IRISA, France



July 12, 2017

It looks like you are trying to do bioinformatics in Excel

Download AskOmics?



Version 1.0

### Outline

#### A nightmare of data

#### Integrating and querying data

- Why it is difficult
- Why RDF and SPARQL are relevant

#### What AskOmics is useful for

- Integrating data
- Querying data



## A nightmare of data















## Function



#### Function



## Genomic/genetic map





#### $\times$ genes $\times$ experiments $\times$ organisms

## Integrating and querying data

#### Integrating and querying data

- Why it is difficult
- Why RDF and SPARQL are relevant

### Data everywhere! (aka death by spreadsheet)

Sie Lat Ven joset famet	24								s_guti.tsv - Li	breOffice O	alc					_ 0 X	
B-B-B-BAS	₩ <u>B</u> le	Edit Vew j	insert F <u>o</u>	mat	Sheet !	lata Jools j	<u>M</u> indow <u>H</u> elp									×	
		• 😕 • 🖃 •	· I 🗟 🖴	6 <u>19</u> 1	¥ 💁	🖏 • 💰 I 🧐	• @ • 1 M	🌾 💷 🖽 🕴	🖶 🎚 I 🖑	34 👬 🛠	। 🔹 🏙 👂	8 😂 🔁	🗉 I 🏦 🖽 •	🖃 l 🥒			
which pairs pairs (according	ere Ubr	ration Sans	• 10		<b>B</b> <i>I</i>	U •   A •	0 ·   E E	a   🗐 🖽	1 1 1	% - 🙏	0.0 🟗   🎎	🐹   🔄 🔄		· 3 ·			at Sheet Data Ta
2 N29 883000 N29 883011	AL		• <i>K</i>	Σ -	DEte	st										<b></b>	k i 🖌 🗈 💼 - 🛷
4 N29 993012 5 N29 993013			Δ.	_		n	c	1	D	1		1 5	1	6	H I	-	
6 7028 200004		DEtest						74 and an east	lana in baardiilaa d					-			
6 P029 00000a	2	ACYPI08849	5 pla pda	Yew y	sart reas.	e Sheet Data 3	ols mindaw their						_	*			= crthodb_gene
9 P029 000009 58 2029 000014	3	ACYPI00771	77 8		1 23 23 15	IN ROT	140-01-146		1.27 14 54 19	0.6.2.0	1 (0 <b>1</b> 1 1 1 1 1	III - 🗆 😿					8
11 2029 000025	4	ACYPI00137	78   Liberati		-	B/U-	A - 0 - = =	- 1 - The local division of the local divisi	E COL B + M	ao 11 14 12		- E - E -				me :	teeserthess group
12 CON 000006	5	ACYPI00778	77.0				<u> </u>			*** UU : 304 .04		u 🖬 121		-			W0000
4 2028 860028	6	ACYPIOD100	// pa		- 74 -	a pres	1 6	1 0		6	6	1 8 1					W1000
18 NOS 000034	2	ACVPIA0173	) 93	test		CONCERNO PORT	15 condition@cand	ten cendton@cer	setion leaf 5		CVALUE .					_	V(0000
17 NOR 880030	1 .	ACYDIODIO	2 40	YIP108048	5-17A 89 0	ACYP1000495-0	5A 20	242	17.5	619570036002	0.0048433509532	23					W0092
19 1028 000000 8201		AC 1 P100485	10 4 AC	YP808777	1.74 std r	ACVPR017774	DA 199	20	-4./2	125101402065	0.007058903518	27				0	W0002
28 NO28 000004	, ,	AC 1P100543	8 40	11908778	TA SU D	ACVPROTTET	tA su	9/1	7.03	\$75800634563	0.0118106747833	U				025	N0092
A2 2028 80000	10	ALYP105338	52 6 AC	YP100100	LRA BU D	ACYPI001004-F	RA BOI	241	-1.05	422599930214	0 8142405918483	13				Jx	AV0002
28 2029 000021	11	ACYPI43396	3-R 1 AC	YPH8173	PA 18 20	ACYPHILTS-R	10	20	5.72	943285953656	0.8262536703863	91	_	0			AV0032
25 N28 102022	12	ACYPI00510	20 2 40	W#+89 YP80543	RA sol o	ACYP1005432-6	tA sol	201	1			D	ي د بيه علي من التوري	Ere24 is Calc			- 0
36 2029 000024	13	ACYPI00599	00 00 AC	YP106338	1.7A 82 8	ACYP1003382-7	NA age	avt	Fie Edt Yes	(event Spread	Mont gate Jools	Madim Hale				-	
28 X028 000025	14	ACYPI00755	54 13 AC	YP143396	PA 56 (9)	ACYP143396-P	a. 60	2/1	) 🗹 - 🐸 - 🖬	- 1 🗳 🚔 (%)	¥% €:-∛!	이 - 이 - 1 🛱 💝 1	<b>T E H E I</b> (7 )	1 (4 %) 📽 🏨 🖓	· 🗶 🛞 🔁 💭 🔐 🛄		
28 7028 080027	15	ACYPI06423	20 10 AL	17100533	NA 63 0	ACTIPIOSSOOP	CA BE	20	Liberation Care		B/U-A	- <u>0</u> -   K X X	🗃 💠 🐮 🛤	🙏 - % oa 🕅 🔛	ﷺ ∉ ∉ ⊡ • ≣ •	- 12 - I	
10 NOS 000028	16	ACYPI00256	10 DE AC	YP104755	LAA SU D	ACYPIOTS547	TA SU	9/1	4.1	· 1. 2 ·	- feriest						
2 N29 10002a	12	ACV/7100403	15 AC	YP106423	ARA BO D	ACYP264239.8	RA sol	9/1				C	0	6	P	0	
8 N29 0000b	11/	AC 1P100421	3 18 AC	YP108258	2 65 AR4	ACYPI002569-F	bA ag	20	2 401712000	DLTA INT INT	AC YTIODOOL FA	CLEAR SOLD COORSES	Canadian and Canadian	10 8143735400	2223 1.45764843795412F-		3 (
1028 800030	10	ACTPIU8271	O at Ar	124142275	LPA IN A	ALTPRIMITOR	the set	20	5 AL 1792001	DD-MA gut gut	AC111020223-RA	24	247.8	-4.5025002262	2834 1.45754843/9541/5-	10	
5 N28 00003e	19	ACYP100328	32 19 AC	YP108328	o be AR-	d ACYP900282-0	IA sol	80	4 AC1793000	15-RA gut gut	AC191000115-RA	24	2/3	-417371879525	5644 3.98586873471807E-	12	2.
8 N25 10000	20	ACYPI00887	1 20 AC	YP108887	RA 59 0	ACVP808871 P	ra su	2/1	5 ACTP12003	NAME AND BUT	ACTIPIOSOSIS KA	25	2073	5 1014 College	1558 4.420735291855828-		
9 2028 000031	21	ACYPI01018	32 21 AC	VPECO18	2.RA bgl g	ACVPIELOSE2-6	A sgl	24	2 AC1795222	SEA out outs	ACYPIG2226-RA	24	2/2	-6.01452904743	7162 6.187190895415420	10	0
BIL-1024 000032	22	ACYPI00352	20 23 40	YP80352	DRA SH D	ACYPIONISS A	04 50 74 50	9/7	8 ACYPI0173	ID-RA gut guta	AC1PI36730-RA	24	202	-60.4044304834	422 7.28920721958800#C-	10	2
2 1029 000034	23	ACYPI00955	58 24 AC	YP108202	RA sa c	ACYP1002024-P	loa Ad	gut	38 46191211	DRA and add	ACTPULLIS PA	80	100	-9.74494319944	LEAT R 350248548003338		1
0.12(2)100025	24	ACYPI00203	25 AC	YP108892	RA sal s	ACYP1008920 F	6A 20	21	31 ACYP93063	165-FA gvl gvb	ACYPIOISSIS PA	24	201	-12.4545283872	2054 1.189930305600022	17	
( ( ) ) 🔶 adabei gene	- 14	A C V DIODEOG	28 40	19124582	LWA SH E	ACVPICERELAN	6A 59	2/	AR AL 1112001	100-KA_9V2_9V2	AL19908852-KA	245	2474	4.3/10/004/90	1283 1228120828954/L-	LT	
K fiel	23	AC 1F100892	28 40	19908/08	FRA 50 0	ACTIPATION A	01 201	9/1	34 ACYP10738	177.RA put put	ACYPIOTISTT-RA	24	9/74	4.3280808744	172 1.834805488482236-	LP	
heet 1 af 1	20	ACTPI06881	28 AU	VPREM.	TRA SU D	ACYMOUS/28	KA SU	9/7	28 461192000	12-HA_0VE_0VE	ACTHORNER	24	264	4.3/32545232	4392 1.034009409402206-	U.	
	27	ACYPI0B317	2 38 AC	19100953	2.RA_59_9	ACTPRODUCES	kA pgi	253	36 ACYPERIO	123-PA and and	AC179081323-RA	25	202	-4 14884140290 2 4884 345343	5600 0.13720955582514C- 1112 0.16101805530000C-	07	
	28	ACYP100700	33 32 40	VPEORE	NA SH D	ACTIPIDINE R	A 50	2/2	28 ACTIVITIE	1/2 PA. BVD. BVD	ACTHIOILET2-RA	2.5	207.0	-4.07704444920	1414 3.589394599911595	ur	
	29	ACYP100357	2 33 AC	YP131076	PA SELOV	ACYPIIL074-R	4 50	9/1	38 4.01793005	18-PA pd p0	AC1PI003518-RA	25	2002	-2.5209370371	1248 3.074887434191738	17	
	30	ACYPI00950	13 NG AC	YP80158	2-RA_22 0	ACYP9001992+	6A 20	212	48 AS 1712/14	ID-HA OUT OUTS	AL TITUTION A	2.5	2074	-5 5918958724	1214 4 223450-50905-C-	V V	
	31	ACYPI30861	LIF MALLY	19103345	NA 59 9	AC17980334544	0A 20	20	22 AC191206	24 RA pvt pvt	AC191000728-RA	2.5	2/2	-219817977838	1547 4. Pelchosabidanas	LP	
	32	ACYPI00884	12 17 AC	YP108508	A pd p	ACYP10060924	tA age	201	28 ACYPROT	28.8A pv1 pv5	AC191037728.RA	24	264	-5.42908401100	1286 A 996599882513448-	17	
	99	ACVPI21074	38 AC	YP808553	RA syl p	ACYP1008932-F	tA sul	201	25 AC1793090	53 PA and and	AC1F9099253-RA	2.5	9/3	5.09652799623	7258 6.741468096819968	17	
	24	ACYPIOD100	39 AC	YP108495	LRA BU S	ACYPI004914-P	DA age Ad	20	25 ACYPHELE	TRA put puta	AC17H8187-RA	24	2/3	-2 50450889626	826 7.364807964283020-	17	
file Edit View	34	ACTPI00195	41 47	YP88352 V2883115	A sol o	ACVPR83528-P 4 ACVPR83528-P	DA and	20	27 AC1197/21	10-PA_B0_80	ACTIVICIES RA	25	202	-2.9050/57/53	NEK2 8.757248190000042-	U7	
D 08 - E	35	AL 1P100345	42 AC	YP100957	T-RA and p	ACYP1009577-8	NA Hel	241	29 ACYT93721	42-8A gut gut	AC171072142-RA	24	2/2	4.1709457298	1072 1.00974898070150E-	16	
· · 👝 · 🖷	36	ACYPI08232	48 AC	YP100148	a los AR-1	/3 ACYPI001431/	tA set	ava	28 AC1792040	12-PA_8VE.8VS	AC1990B111-RA	245	262	4 82211029904	1991 1.289923/51/28256-	10	
L	37	ACYPI00509	92	n <b>e</b> r	C agl va a	160		_	31 AC1791000	13.84 pd pd	AC1P1082017-KA	25	202	2 20313133021	7254 1.431408728062078- 5551 1.685378232627628-	10 16	
Liberation San	15 38	ACYPI00593	32	1		-	- A Rodal City	AND DESCRIPTION	AK AC1794201	DHA gut guta	AC1714L002-MA	24	247.8	-2.14894811/90	0000 1.0003/823242/02L	10	
	39	ACYPI00491	4	-	_		Date		M ACYFRIG	WARA and and	ACYPIOROPEZ-RA	2.1	2/3	-417415900970	H14 1.68537623242762E-	N	
u.	40	ACYPI08352	28 707 39		ACTIN	0332070A 2	Deta	1993	28 46111210	10 11A BVL BVL	ACTHOR/BELKA	20	202	-3.899976280	1200 1.9218/2013/095/00-	15	
0	41	ACYPI00312	R-RA so	1 auti	ACYPIC	03138-RA	1	auti	22 AC1793090	07.0A ad ad	ACVENNESSOF RA	25	202	1.45431.099313	2875 1.92197361339536E	6	
~	42	ACYPI00951	77.PA 40	L outi	ACYPIC	09577.PA	1	auti	38 AC179304	115-PA get get	AC171004815-RA	24	2014	-3 68665437308	1302 2.41783860643818C-	16	
condition	0	ACYDIOD140	1 0.4 49	d and	ACYDIC	01401 00		2003	40 AC191214	TRA DE DAS	ACTPIZHIST RA	24	1072	-2.68702057523	2075 2 69005443548096	10	
2 aut	a 43	Pro 1 = 100145	11:NA 50	1 440		AT#AT-KW 21	0	160	41 AC 171264	IN A DID DIA	ACYPI26400-PA	24	264	2.42131.730333	388 2.737523076748486-	16	
2	1 1 1	N.M.	DE col y	or out	-				43 4/ 15101	ALLON AND AND	ALTINA/20-RA	20	202	2 92509992560	HER ZINZZOWO/OULBER-	10	
a guta	9		0.C_301_4	.s_400					-				1				- 10°
4 501	s 🕺	Find					Rnd All	matted Displa		re for a fo		And ALC Residence	Inclust we have	and the second second			
5 902	C Char						Def	- 44	- 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1		200	Benada	a contrary. Barry Ca				
- XXX	Shee	R 1 0F 1					Den	NHL	Sheet 3 of 3			Cefault		× 6	Annape ; Sarv D	-	+ 1639

#### Definition: Entity

Anything that can be identified (i.e. the things we can talk about)

- some informations about an entity in a repository
- other informations about the same entity in another repository

your question requires to combine entity descriptions from multiple datasets

Only possible if:

- Entities are identified
- Datasets use the same identifier to describe the same entity

Good luck with your spreadsheets! :-)

#### Data description involves hierarchies

- Data are precise
- Queries involve more general criteria

your question requires some reasoning (often based on hierarchies and ontologies) in order to reconcile the data and the criteria

Only possible if:

- Knowledge has been formalized (e.g. in ontologies)
- Query engines (more or less) gracefully handle
  - linking data and ontologies
  - reasoning on the ontologies

- Identify entities (uniformly) so that multiple repositories use the same identifier when they refer to the same entity
- Describe entities
  - their characteristics
  - their relations with other entities
  - both can be scattered in multiple repositories
- Query descriptions even if scattered in multiple repositories
- Perform reasoning (based on domain knowledge) over these descriptions

These points exceed the "classical" relational model's capabilities

#### The good news (1/3)

Semantic Web technologies (RDF, SPARQL, OWL) address all of these

O. Dameron

## Linked Open Data (in 2009)



## Linked Open Data (in 2014)



## Linked Open Data

Semantic Web technologies

April 2016: (according to http://stats.lod2.eu)

- 149.10<sup>9</sup> triples
- distributed over 9960 knowledge graphs
- A treasure at your fingertips (or a nightmare of spreadsheets)

# Linked data are here... but still have to be adopted by end users

"Real" users

- do not contribute (yet) their data to the LOD cloud
- do not use the LOD cloud for analyzing their own data (yet)

## RDF principles

Entities and relations are identified by their URI RDF dataset = directed graph of triples

- subject: the entity we are talking about
- predicate: the relation used to describe the entity
- **object:** (one of the) relation's value for the subject

Example:

Alice plays violin Alice plays guitar Alice knows Bob

#### The good news (2/3)

Don't worry about RDF, AskOmics generates it from your csv

## SPARQL principles

SPARQL = query language similar to SQL Variable names start with a question mark What instruments does Alice play?

```
SELECT ?instr
```

```
2 WHERE {
```

}

```
p1:Alice mus:plays ?instr .
```

```
3
4
```

1

Returns:

violin

guitar

#### The good news (3/3)

Don't worry about SPARQL queries, AskOmics generates them for you (and runs them as well)

## What AskOmics is useful for

- 3 What AskOmics is useful for
  - Integrating data
  - Querying data

### AskOmics

#### Integrating data

- Import your data files
  - CSV or TSV
  - RDF
  - GFF
- Import public knowledge bases (GO, Reactome, NCBI taxon,...)
- Declare (remote) SPARQL endpoints (in progress)

#### Querying data

### AskOmics

#### Integrating data

#### Querying data

- Graph-based user-friendly SPARQL query composer
  - based on an abstraction of your data
    - depends on the data structure (small)
    - not on the data themselves (possibly huge)
  - can be simplified (hide a portion of the graph)
  - can be enriched (shortcuts and virtual links)
  - modular design (select/deselect datasets)
- Span multiple SPARQL endpoints (in progress)
- You do not have to see the SPARQL code
- Save the query result (obviously)
- Save or import SPARQL query

## Installing AskOmics

• Github AskOmics

https://github.com/askomics/askomics

- Instructions
  - See https://github.com/askomics/askomics/wiki
  - \*nix: docker-compose
  - Windows and \*nix: virtualbox (but performance limitations)

#### AskOmics demo



Based on Anthony Bretaudeau's demo on aphids data at the 2017 Arthropod genomics symposium - Chicago

## Acknowledgments

- Meziane Aite
- Arnaud Belcour
- Charles Bettembourg
- Yvanne Chaussin
- Aurélie Évrard
- Xavier Garnier
- Maël Kerbiriou
- Colleagues from Nantes for their insight on federated queries
  - Pascal Molli
  - Patricia Serrano Alvarado
  - Hala Skaf
- Sylvaine Bitteur (INRA / Agrocampus Ouest) for the logo