PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

We don’t talk about Semantic Web Technologies

ACS Fall 2023
13th August 2023

Dr Samantha Kanza
University of Southampton

https://www.psdi.ac.uk/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

About Me

- Senior Enterprise Fellow at University of Southampton
- Pathfinder Lead & Researcher for PSDI Project: Process Recording
- Coordinates AI4SD & Future Blood Testing Network
- Research Interests: Semantic Web Technologies, IoT, Research Data Management, Digitisation, Lab of the Future, Paperless Labs, Re-use of Technology
- @SamiKanza

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Lets talk about the Semantic Web

Conclusions

- **The Semantic Web is coming!**
  - Joint development between DARPA/EU/and W3C communities
  - Languages and tools are available to play with
    - Http://www.daml.org/
  - W3C interest group available for those wishing to join the discussion
    - Www-rdf-logic@w3c.org (live or archived)
  - Ongoing DoD and commercial projects
- **Come join us**
  - Submit ontologies/marked up pages
  - Develop tools or help test ours
- **Get in on the next big thing early!**

www.daml.org

https://www.wi-consortium.org/wicweb/pdf/wi-hendler.pdf

WI-2001: Maebashi, Japan

MiND

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Common Misconceptions

The concept of machine-understandable documents does not imply some magical artificial intelligence allowing machines to comprehend human mumblings. It relies solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data. So, instead of asking machines to understand people's language, the new technology, like the old, involves asking people to make some extra effort, in repayment for which they get major new functionality – just as the extra effort of producing HTML mark-up is outweighed by the benefit of having content searchable on the web.

Berners-Lee, T. and Hendler, J., 2001. Publishing on the semantic web. *Nature*, *410*(6832), pp.1023-1024.

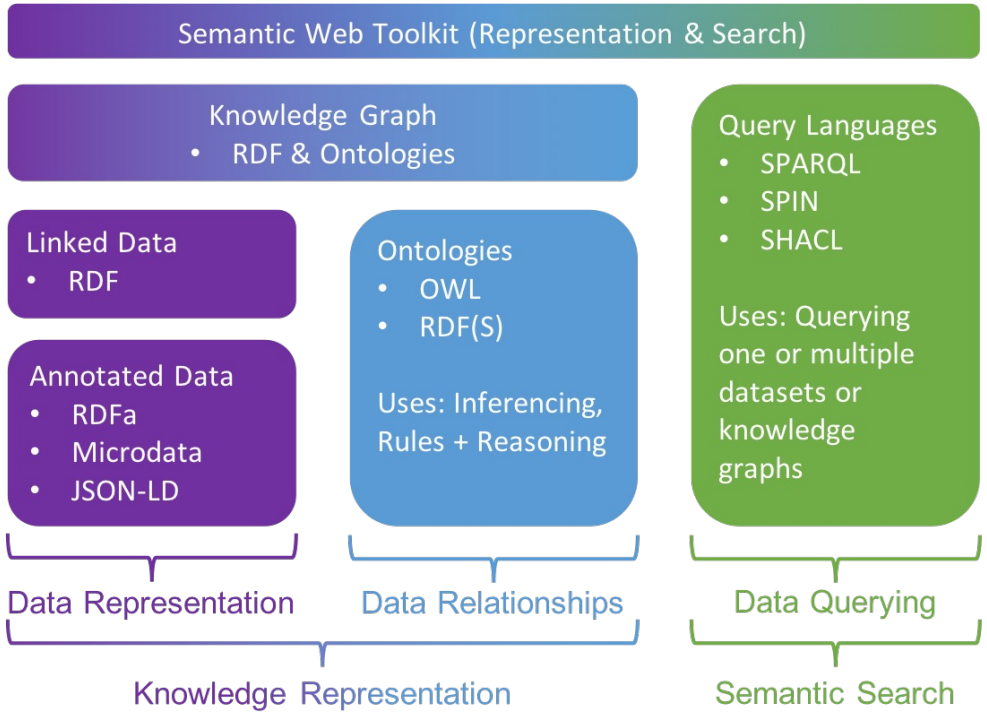
Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The semantic web. Scientific american, 284(5), pp.34-43.

The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings.

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

So what is the Semantic Web?

- The Web of Linked Data
- Creating machine-readable/understandable data
- A way to bring context and meaning to data
- A set of common standards for data representation, integration, and search



https://www.sciencedirect.com/science/article/pii/B978012801238311520X?via%3Dihub

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

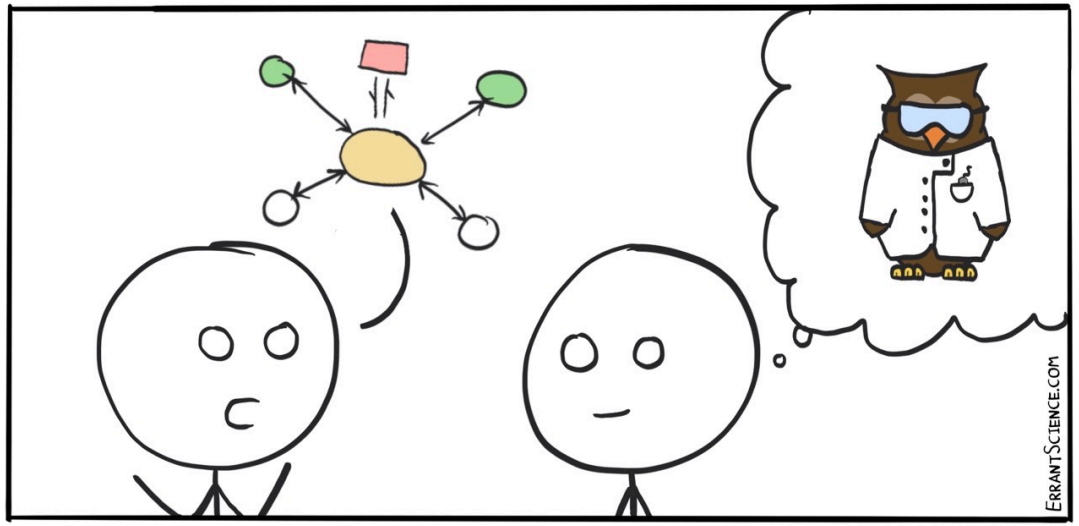
So where did this all go wrong?

https://www.pinterest.co.uk/pin/539306124105554567/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Ontologies

- Too many ontologies (and yet not enough?)
- “Lets just make a new one” mentality
- Lack of standards for Ontologies (aside from the biology sphere)
- Ontology projects frequently don’t consider the full data lifecycle
- Ontologies aren’t always well maintained
- Ontologies aren’t always FAIR
- Ontology tools are poor or expensive



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Linked Data Creation/Conversion

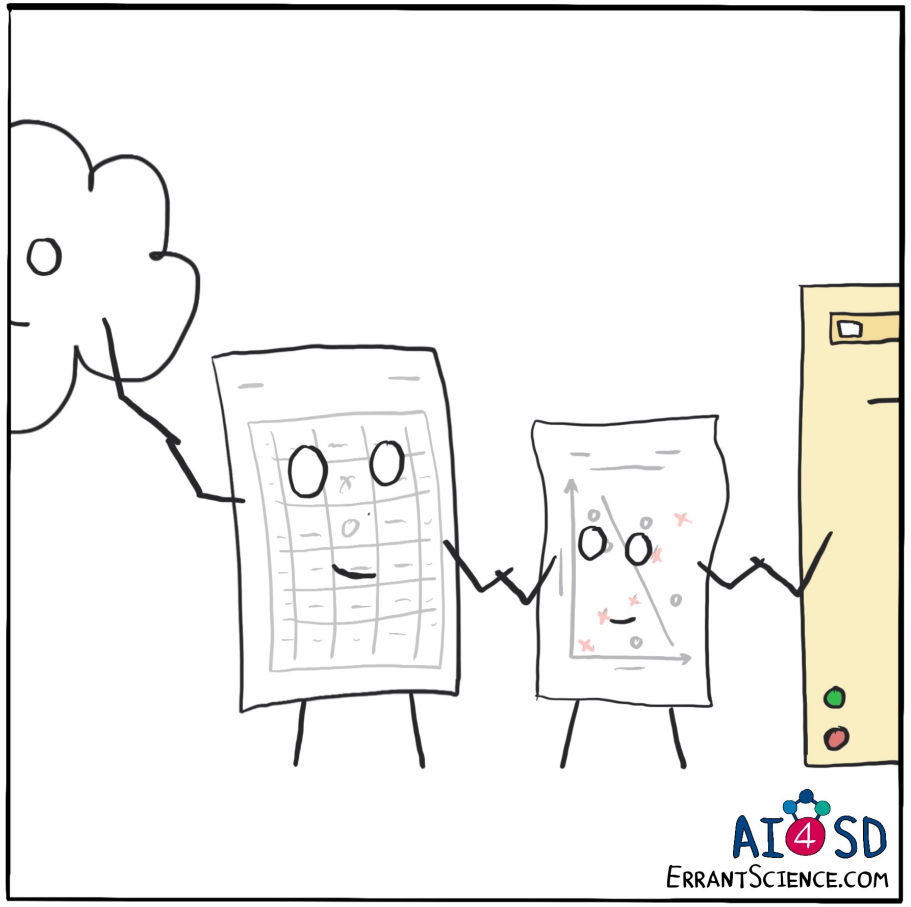
- Tools to convert structured data into linked data are still lacking
- R2RML even with additional libraries is repetitive
- Existing tools (OpenRefine or Karma) are very clunky and require a lot of human editing
- Writing custom scripts seems to be the best current method but isn't very user friendly

https://imgflip.com/memegenerator/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Data Storage & Data Types

- Some data doesn't lend itself well to being represented semantically (e.g. numerical classifications)
- Large datasets lead to even larger knowledge graphs
- Issues with reading, writing and deleting data with certain (free) versions of triple stores



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Given these issues, why should we care?

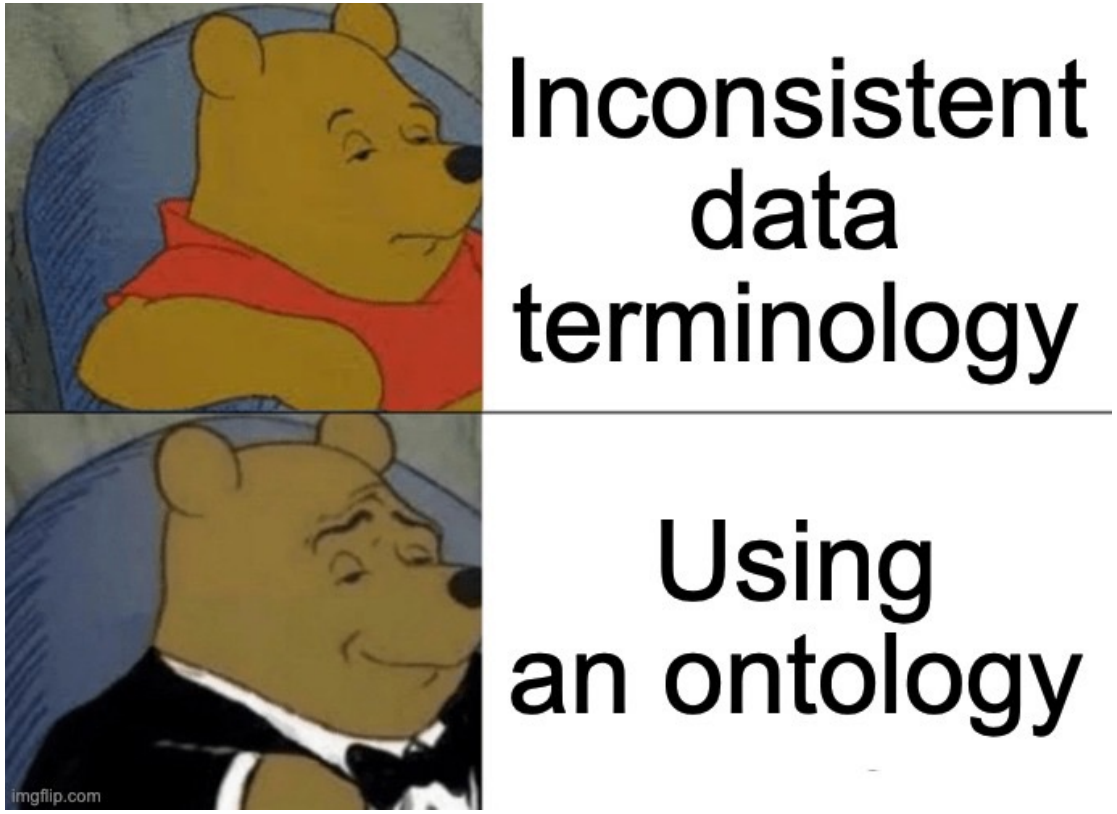
- Defining Common Shared Vocabularies
- Machine Readable Data
- Interoperable Metadata
- Linking Datasets
- Inferencing
- Semantic Search
- Unlocking the potential of AI/ML

https://imgflip.com/i/4x69cr

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Common Shared Vocabularies

- Define a common set of terms to describe scientific data concepts consistently
- Markup your datasets to talk about the SAME concepts

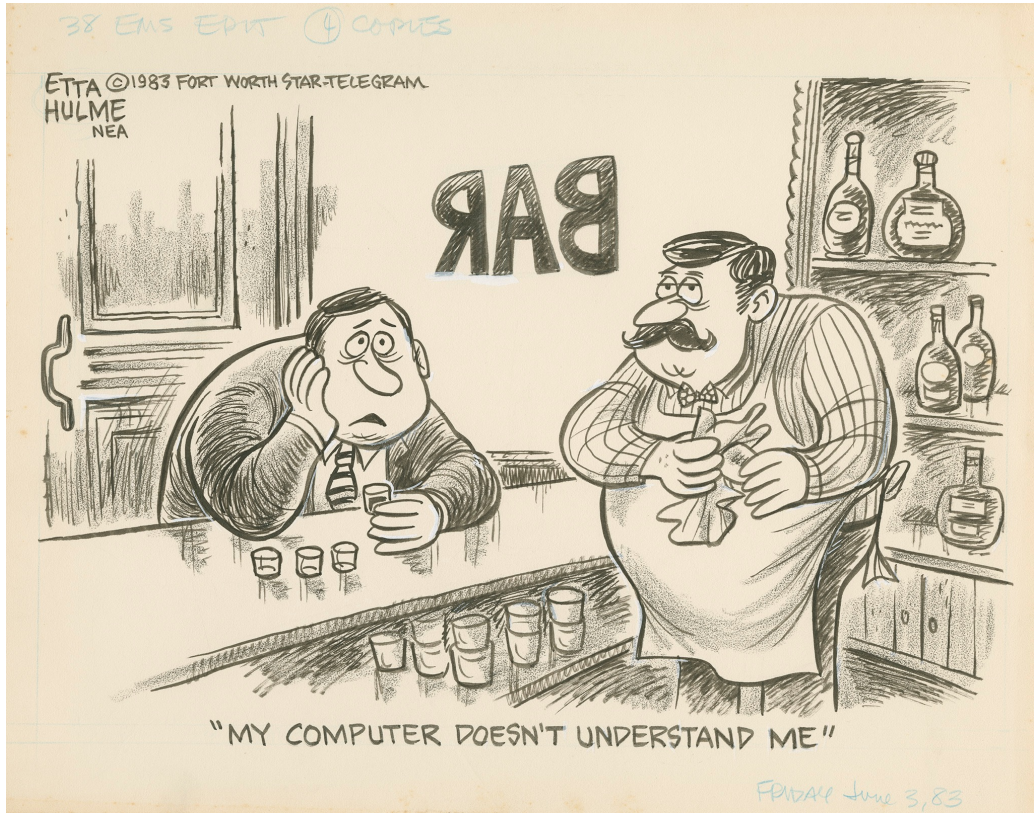


https://imgflip.com/i/7uq27y

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Machine Readable Data

- How can computers be expected to analyse and make decisions on scientific data they don't understand?



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Rich Interoperable Metadata

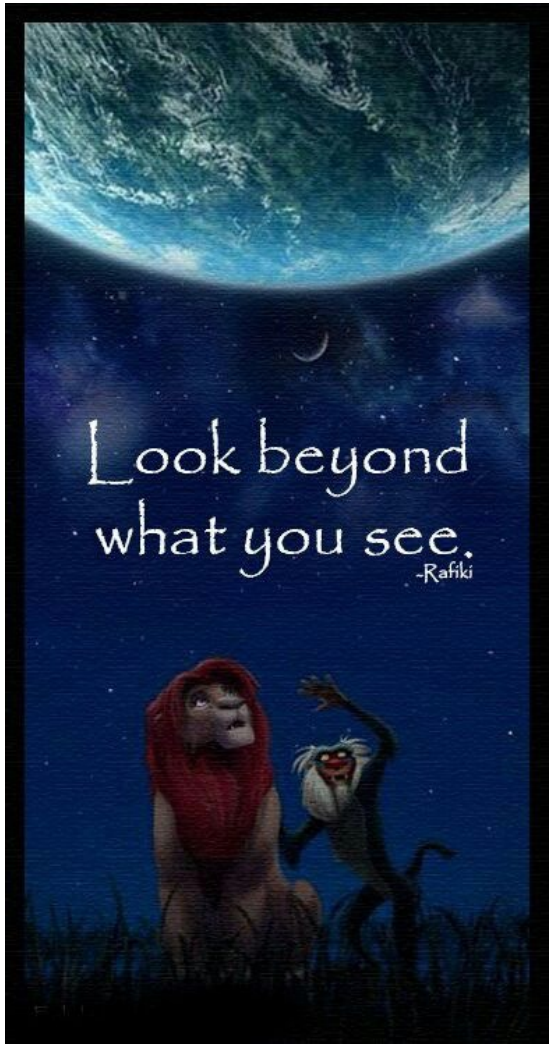
- Adding rich interoperable semantic metadata to your documents / webpages adds a whole new dimension to your data

https://www.pinterest.co.uk/jaci_mize/metadata/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Semantic Knowledgebases

- Scientific Research can be significantly enhanced by linking datasets together to create knowledgebases:
  - Find undiscovered links
  - Answer questions that cannot be addressed with a single data source
  - Very useful for domains like drug discovery which require the integration of disparate datasets



https://www.pinterest.co.uk/pin/2674081014251214/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Inferencing

- Description Logic can be embedded into Ontologies
- This enables machines to “infer” additional information that is not explicitly defined in the data

https://quoteswell.com/cheshire-cat-quotes/

Example: If we know that:

```
Sami isAllergicTo Juniper
Gin hasBotanical Junpier
```

Then we can infer: `Sami isAllergicTo Gin`

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Semantic Search

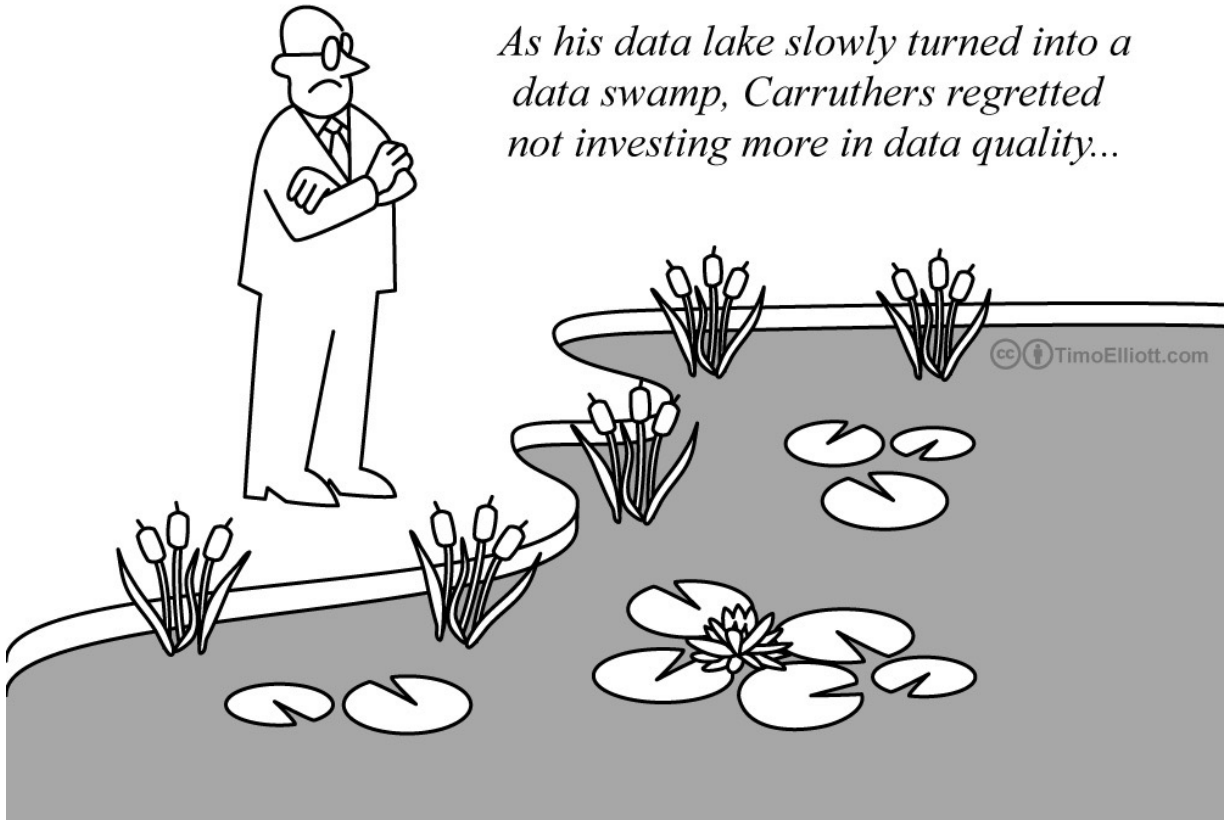
- Search on CONCEPTS across LINKED DATASETS

https://garfield.com/comic?keywords=Jon&page=1113

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Unlock the Potential of AI & ML

- Don’t waste your algorithms time with poorly formed inconsistent data



https://timoelliott.com/blog/cartoons/more-analytics-cartoons

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

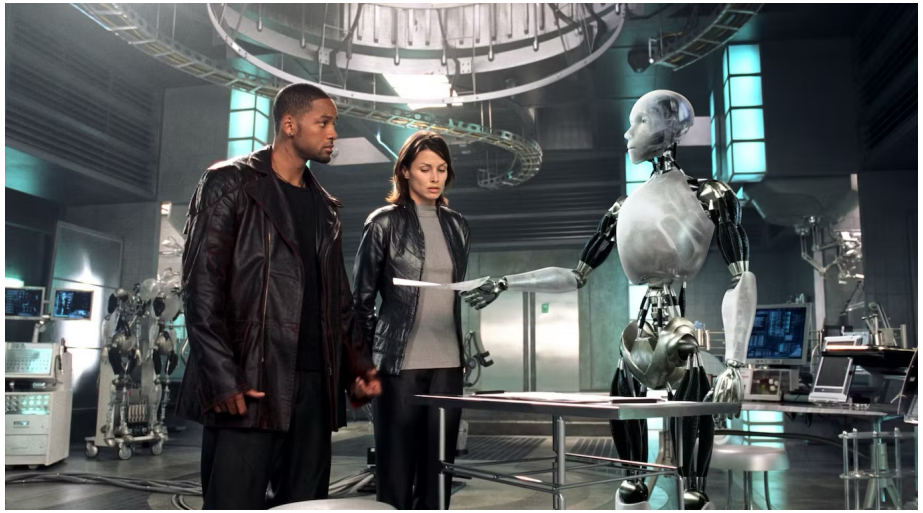
So what can we do?

https://memegenerator.net/instance/68158697/hermione-leviosa-swish-flick

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

You need to ask the right questions

- What is your use case for semantics?
- Do you need an ontology?
- Do you need a knowledge graph?
- How much data do you need to represent semantically?
- What data would benefit from being in a semantic format?
- What do you want your data to look like and how are you going to use it?



https://www.inverse.com/culture/sci-fi-movies-january-2023-hbo-max-i-robot

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

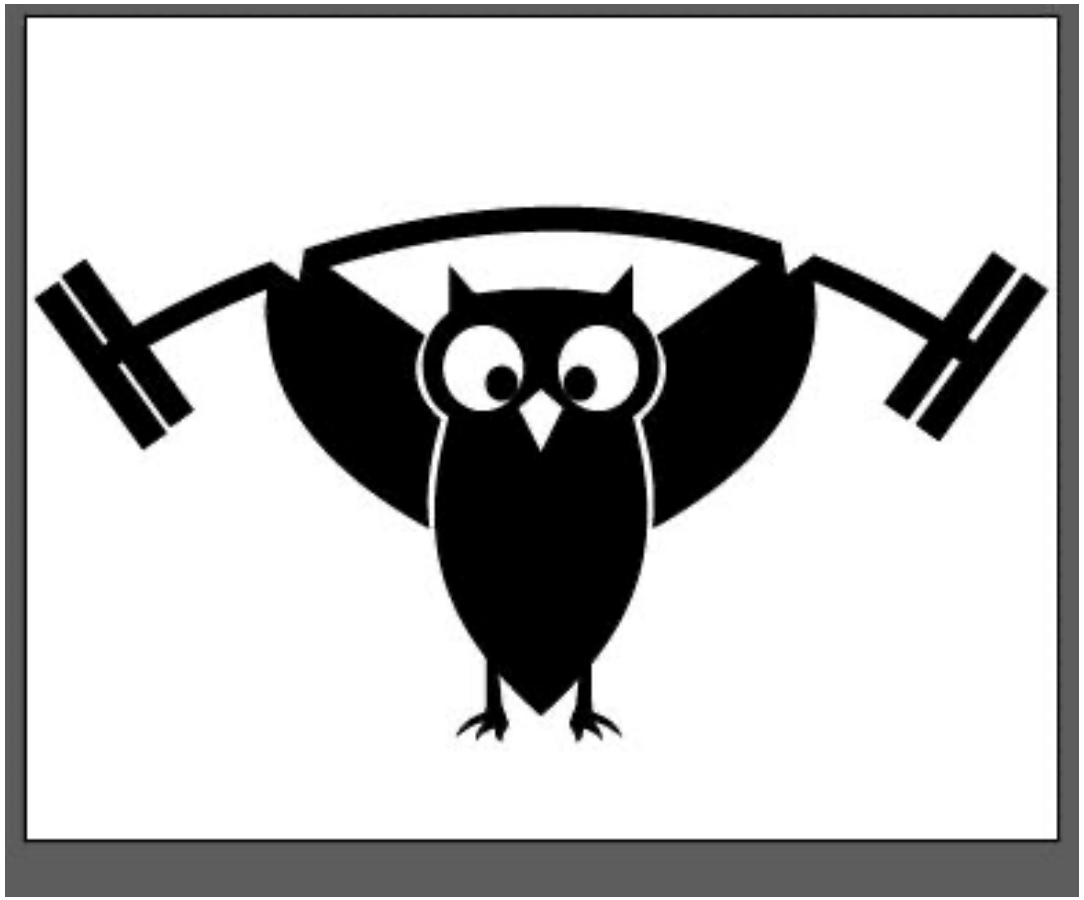
Consider your Conversion in Advance

- Converting data into linked data is non-trivial
- Conceptualising an Ontology vs using it in a dataset are two different things
- You need to consider the entire semantic data lifecycle of a project

https://www.woorank.com/en/blog/implement-semantic-web

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Think about the WHY and the WHAT



https://cuk00jan.wordpress.com/2014/02/14/who-is-alpha-owl/

- Think about WHY you are making Ontologies and WHAT they are for
- Play to OWLS Strengths

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Re-use and extend Ontologies where possible

- There are lots of ontologies out there! See if there's any you can re-use before making a new one (Search Papers/OLS/Google)
- Re-use doesn't just mean ontologies! You can re-use design patterns as well!
- Consider trying to create an extension to an existing ontology if appropriate

https://www.fosteropenscience.eu/content/cartoonreusable-data

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Modularize your Ontologies

- Break your ontologies into smaller related modules

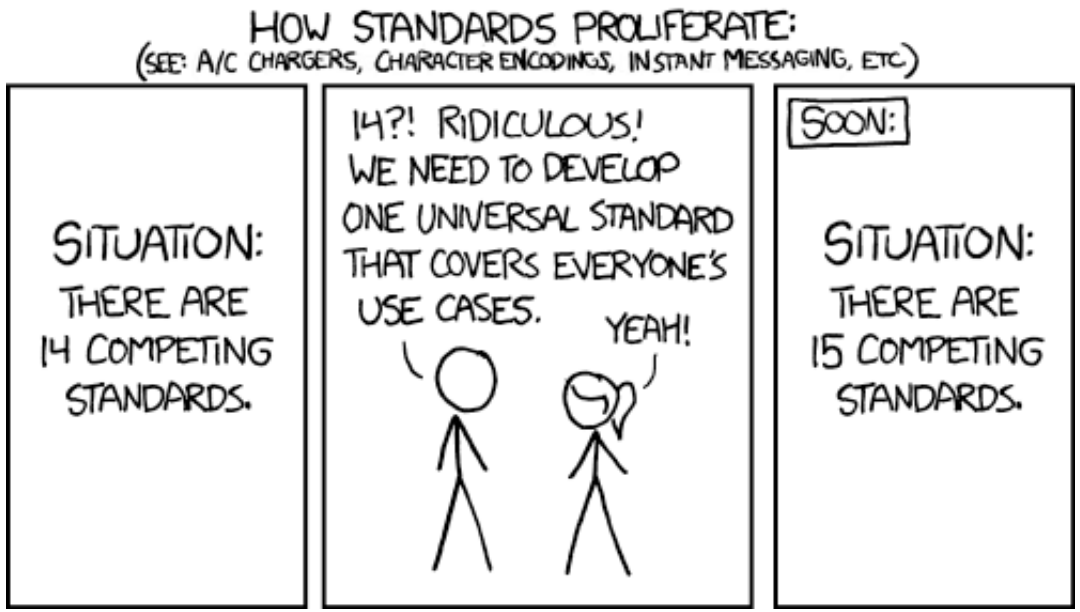
Relation to time / Granularity	Continuant				Occurrent
	Independent			Dependent	
Complex of organisms	Family, community, deme, population		Environment	Population phenotype	Population process
Organ and organism	Organism (NCBI Taxonomy)	(FMA, CARO)		Organ function (FMP, CPRO)	Phenotypic Quality (PaTO)	Biological process (GO)
Cell and cellular component	Cell (CL)	Cell component (FMA, GO)		Cellular function (GO)		
Molecule	Molecule (ChEBI, SO, RnaO, PrO)			Molecular function (GO)		Molecular process (GO)

https://www.researchgate.net/publication/268412593_Ontology-Based_Knowledge_Representation_of_Experiment_Metadata_in_Biological_Data_Mining/figures?lo=1

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

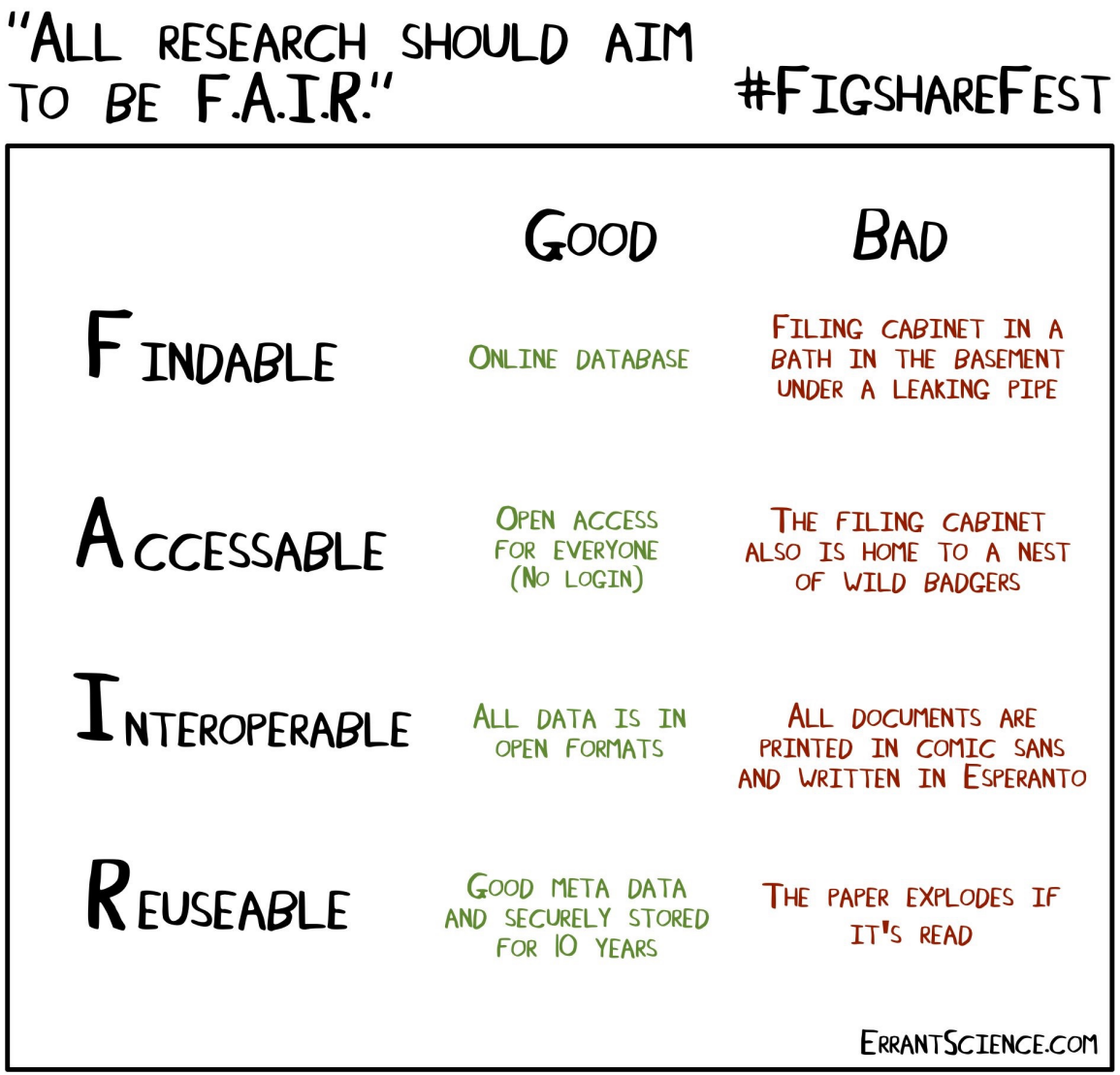
Don’t forget your standards

- Even your standards should have standards
- Consider:
  - Design patterns
  - Upper Level Ontologies
  - Consistency



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

FAIR is a Four Letter Word



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Garbage In = Garbage Out

- GIGO can still apply even in the world of semantics

https://ffineis.github.io/blog/2017/09/13/case_for_big_govt.html

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Conclusions

- We're all in this together and we need to work together!
- A lot of the technology is there, we just need to use it properly!
- This is as much a human endeavor as a technical one!

https://www.thedesigngym.com/skip-collaboration-collaborations-sake/

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Relevant Publications & Talks

- Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J.G., Erjavec, J., Zupančič, K., Hren, M. and Kovač, K., 2017. Electronic lab notebooks: can they replace paper?. Journal of cheminformatics, 9(1), pp.1-15. https://doi.org/10.1186/s13321-017-0221-3
- Kanza, S., Stolz, A., Hepp, M. and Simperl, E., 2018. What does an ontology engineering community look like? A systematic analysis of the schema. org community. In The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15 (pp. 335-350). Springer International Publishing. https://doi.org/10.1007/978-3-319-93417-4_22
- Kanza, S., 2018. What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? (Doctoral dissertation, University of Southampton). https://eprints.soton.ac.uk/421045/
- Kanza, S., Gibbins, N. and Frey, J.G., 2019. Too many tags spoil the metadata: investigating the knowledge management of scientific research with semantic web technologies. Journal of cheminformatics, 11(1), p.23. https://doi.org/10.1186/s13321-019-0345-8
- Kanza, S. and Frey, J.G., 2019. A new wave of innovation in Semantic web tools for drug discovery. Expert Opinion on Drug Discovery, 14(5), pp.433-444. https://doi.org/10.1186/s13321-019-0345-8
- Kanza, S. and Frey, J.G., 2020. Semantic technologies in drug discovery—potential, practical, possibilities. http://dx.doi.org/10.1016/B978-0-12-801238-3.11520-X
- Kanza, S., 2021. AI3SD Video: Semantic Web in Scientific Research–Possibilities & Practices. http://dx.doi.org/10.5258/SOTON/P0078
- Kanza, S., Willoughby, C., Bird, C.L. and Frey, J.G., 2022. eScience Infrastructures in Physical Chemistry. Annual Review of Physical Chemistry, 73, pp.97-116. https://doi.org/10.1146/annurev-physchem-082120-041521
- Kanza, S., Willoughby, C., Knight, N.J., Bird, C.L., Frey, J.G. and Coles, S.J., 2023. Digital research environments: a requirements analysis. *Digital Discovery*. https://doi.org/10.1039/D2DD00121G

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Acknowledgements

- **PhD Supervisors:** Jeremy Frey & Nicholas Gibbins *(University of Southampton)*
- **Electronic Lab Notebook Research:** Cerys Willoughby & Nicholas Gibbins & Richard Whitby & Jeremy Frey *(University of Southampton)* Jana Erjavec, Klemen Zupančič, Matjaz Hren & Kristina Kovač *(SciNote)*
- **PSDI Team:** Simon Coles, Jeremy Frey, Nicola Knight, Cerys Willoughby & Colin Bird *(University of Southampton)*, Juan Bicarregui, Barbara Montanari, Brian Matthews, Vasily Bunakov *(Science & Technologies Facilities Council)*
- **Semantic Web Research Teams**
  - **SEED:** Pistoia Alliance, Pfizer, Biovia, Dotmatics, Sanofi, BenchSci, AstraZeneca, Elsevier, GSK, Bayer, Merck, SciByte, University of Southampton, Bristol-Myers Squibb, ChemAxon, IDBS, Linguamatics, Arxspan
  - **BSP:** Hugo Mills & Hannah Wickenden & Derek Scuffell (Syngenta), Jeremy Frey & Nicholas Gibbins *(University of Southampton)*
  - **S3W:** Susan Halford & Faranak Hardcastle *(University of Bristol)*, Nicholas Gibbins & Mark Weal *(University of Southampton)*, Cathy Pope *(University of Oxford)*

PSDI & Personal Details - Questions

www.psdi.ac.uk

@PSDI_UK

@PSDI_UK

linkedin.com/company/psdiuk

linktr.ee/samanthakanza

Mailing List: *https://www.jiscmail.ac.uk/PSDI*