The FAIR data scientist

Dr Rebecca Lange
Curtin Institute for Computation

WAGUL Research Forum - 23 July 2019
What do Astronomy, Art Conservation and Smart Cities have in common?
one data scientist
I loved astronomy ever since I can remember ❤️⭐

But I also like to build things and study old things. 🔬🎨

And how will all the technological advances change how we live?🤖🏙
● Lived and studied in **3** countries
● Visited **11** countries for work
● Collaborators across the globe
My Journey
Art Conservation

Imaging & Sensing for Archaeology, Art History & Conservation[1]
Astronomy

Galaxy And Mass Assembly survey[^2]
Data Science

Curtin Institute for Computation

Multi-modal analysis - Shiny Web App

RENeW Nexus

Precinct-Scale Data Collection

WG by LandCorp
White Gum Valley, Fremantle

RAC Pulse of Perth

08:36 am
24th Oct 2017

Total load: 815.1 MW
Residential: 204.6 MW
Commercial: 128.8 MW
Industrial: 277.4 MW
Customer: 10 Chapel St, North

Top 6 bus stops:
- Sheep Station, Fremantle
- Stirling St, Fremantle
- Victoria Quay, Fremantle
- South Fremantle, Fremantle
- Burswood, Fremantle
- Woodford St, Fremantle

Top 6 train stations:
- Fremantle Train Station
- Victoria Quay Train Station
- Stirling St Train Station
- Burswood Train Station
- South Fremantle Train Station
How do you share your data?
How do you share your data?

FAIR-ly.
To be Findable:
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:
A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:
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.
The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. [8]
I.1 - exchange of data
We worked with a software developer to write the operational software for our instrument. After lengthy discussions we agreed to save the data as a simple CSV file.

JSON[^11] (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. ➡ Easy to automate with API.
I.1 & I.2 - exchange & vocabularies

FITS[^12] is used for the transport, analysis, and archival storage of scientific data sets (open standard, 1981)

- Multi-dimensional arrays: 1D, images, 3D+ cubes
- Tables containing rows and columns of data
- Header keywords provide descriptive information about the content
  - Agreed standard for e.g. telescope images

---

**Findable**

**Accessible**

**Interoperable**

**Reusable**
I.3 - linked data

The Centre de Données astronomiques de Strasbourg (CDS\textsuperscript{[13]}) provides a service that links various data sources making discovery easy.
The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. [8]
For any of our value-add catalogues for GAMA\(^2\) we followed strict guidelines of how to collate the data and metadata making sure the catalogues and tables had explanations and descriptions detailed enough for any new users to jump right in, e.g.:

- Who, when, what?
- Scope and limitations of data

**R1 & R1.2 - usability and history of data**

Findable

Accessible

Interoperable

Reusable
Most telescope data is propriety for a short period of time before it is being made public, e.g. Hubble Space telescope data is made public after 1 year. Many large surveys work on value-add catalogues which are made public after a period of time or after a journal publication.
"The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole. Many projects and data centres worldwide are working towards this goal. The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible. It also acts as a focus for VO aspirations, a framework for discussing and sharing VO ideas and technology, and body for promoting and publicising the VO." [15]
Projects mentioned
[1] Imaging & Sensing for Archaeology, Art History & Conservation (ISAAC)  
https://www.ntu.ac.uk/research/groups-and-centres/groups/imaging-sensing-for-archaeology-art-history-and-conservation
http://www.gama-survey.org/
[3] Curtin Institute for Computation  
https://computation.curtin.edu.au/
https://shiny.computation.org.au/MMAv0.2/
[5] RAC pulse of Perth  
[6] RENEW NEXUS  

FAIR principle

FAIR examples
https://json.org/
[12] FITS  
https://fits.gsfc.nasa.gov/
[13] Centre de Données astronomiques de Strasbourg (CDS)  
http://cds.u-strasbg.fr/
[14] Barbara A. Mikulski Archive for Space Telescopes (MAST)  
http://archive.stsci.edu/
[15] International Virtual Observatory Alliance  
http://www.ivoa.net/