A tool for the repeatable generation, and automated documentation, of Land-use Classification Maps

Ben Jolly, Markus Müller, David Medyckyj-Scott, Raphael Spiekermann, and Anne-Gaelle Ausseil

LANDCARE RESEARCH
MANAAKI WHENUA

What am I talking about?

- (Very) quick data provenance 101
- pyluc – what it is, why we made it, how it works

What is data provenance (data)?

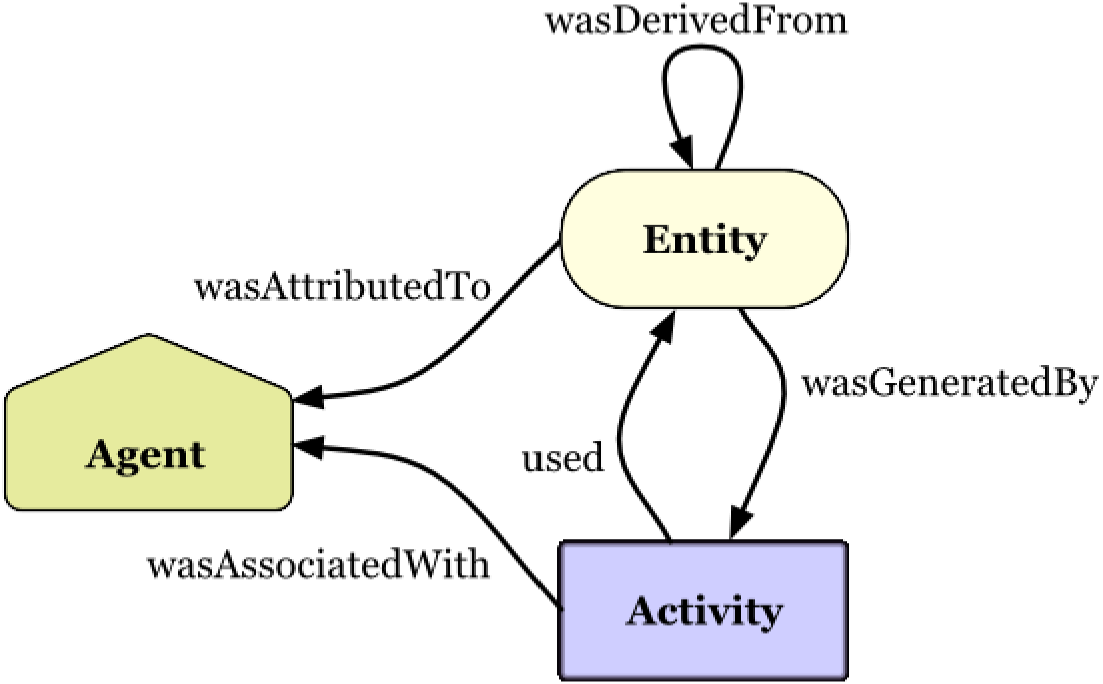
- **Part** of the **metadata** surrounding a dataset
- A record of ***what*** has happened to some data, ***where*** it happened, ***when*** it happened, ***how*** it happened, ***who*** did it, using ***which*** tools/instruments, for what purpose (***why***)

W7 model

What does it look like?

- Can take the form of
  - A file (JSON/XML/PROV-N/etc.)
  - A hosted service/site with interactive visualisations (ProvStore)
  - A series of blockchain transactions

The W3C PROV model



source: W3C PROV-PRIMER

Why should I bother?

- Data quality
- Audit trail
- Attribution
- Informational

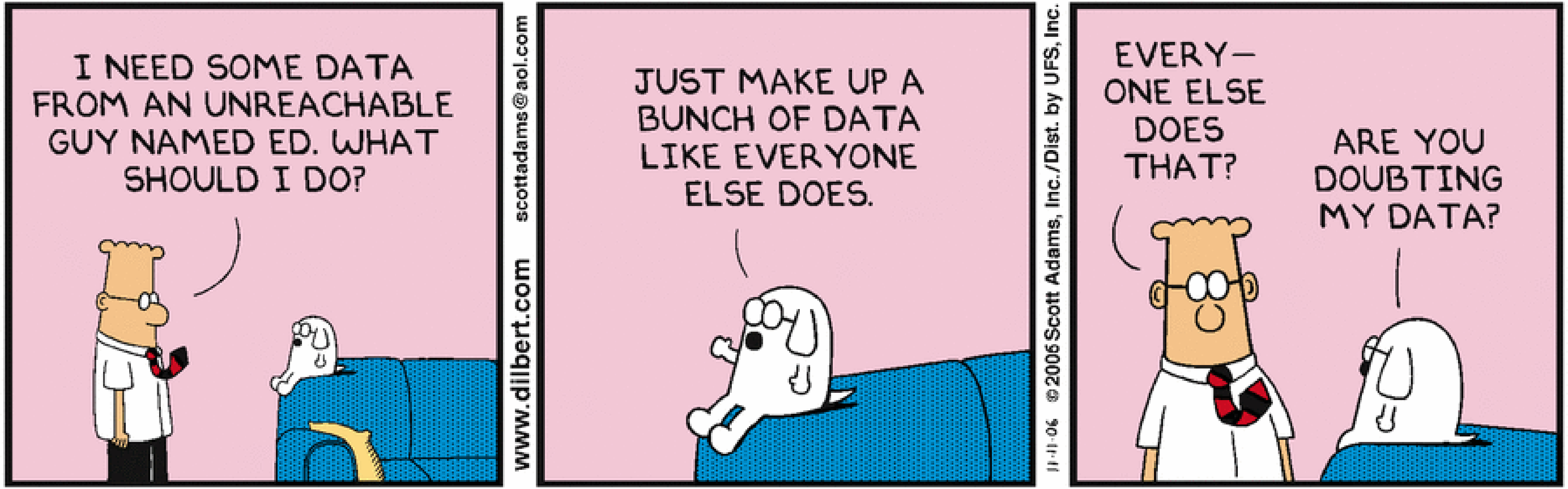
People are going to start asking for it
(UK gov. is leading the charge on this)

What is pyluc?

- A scalable Python framework that ingests scripts and produces geospatial datasets accompanied by provenance data and technical documentation
- A single script defines a dataset, anyone can use it to reproduce results and documentation

Why pyluc?

- Open, reproducible research ***more easily***


I NEED SOME DATA FROM AN UNREACHABLE GUY NAMED ED. WHAT SHOULD I DO?
scottadams@aol.com
www.dilbert.com
JUST MAKE UP A BUNCH OF DATA LIKE EVERYONE ELSE DOES.
11-11-06 © 2006 Scott Adams, Inc./Dist. by UFS, Inc.
EVERY-ONE ELSE DOES THAT?
ARE YOU DOUBTING MY DATA?

Why pyluc?

- Technical documentation and provenance data can be annoying to create manually, difficult to keep up to date with regular changes to methodology
- Existing processing methods were not scalable

How does pyluc work?

1. Initialisation
   - Ingest definition script, initialise framework
2. Data marshalling
   - Request, download, extract
3. Parallel, tile-based, data processing
   - Rasterise/re-project, apply logic
4. Clean-up
   - Merge tiles, vectorise
5. Documentation
   - Code introspection, recording internal links

How does pyluc work on Pan?

- Single SLURM job, starts marshalling then:
  - Resubmits itself with cooldown if waiting for data
  - Spawns array job(s) for tile-based processing
  - Spawns clean-up/doc job dependent on array jobs
- All this from one SLURM script (no templates)
- RAM disks for staging (heavy I/O on Pan)

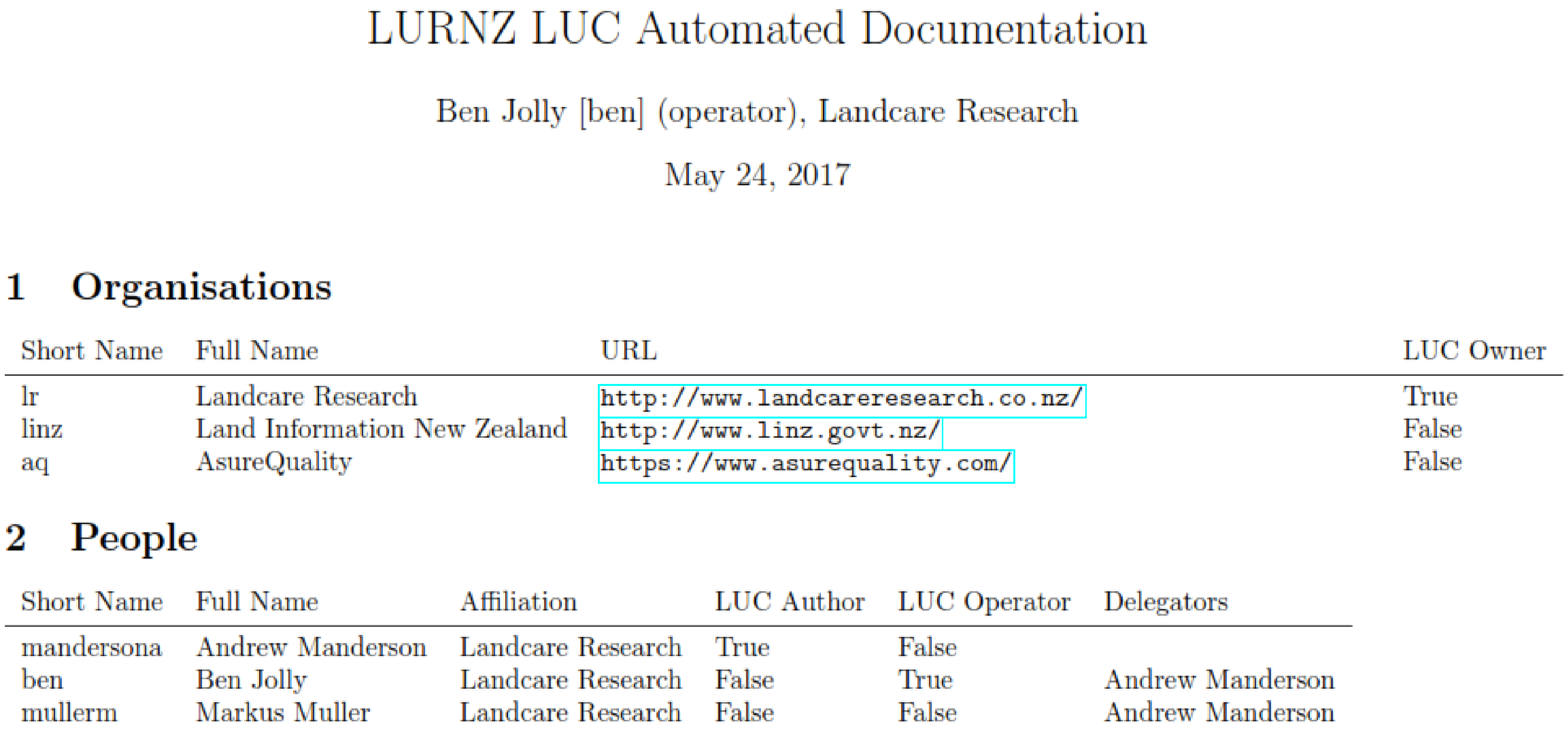
So, what’s in the definition script?

- Basic metadata
  - Name, author(s), extent, resolution
- URLs to input data sources
  - Currently LRIS-only, other Koordinates sites soon
- Logic to be applied to that input data
  - LUTs, Python functions (anything goes)

And what do we get?

- A *.kea (raster) file for each logic step
- An optional *.shp (vector) file for the final step
- A *.tex file for human-readable documentation
  - Authors, input data, relevant organisations
  - Logic steps and how they relate to one-another
  - Syntax-highlighted code snippets describing logic
- A *.provn file, optionally uploaded to ProvStore

Documentation



LURNZ LUC Automated Documentation

Ben Jolly [ben] (operator), Landcare Research

May 24, 2017

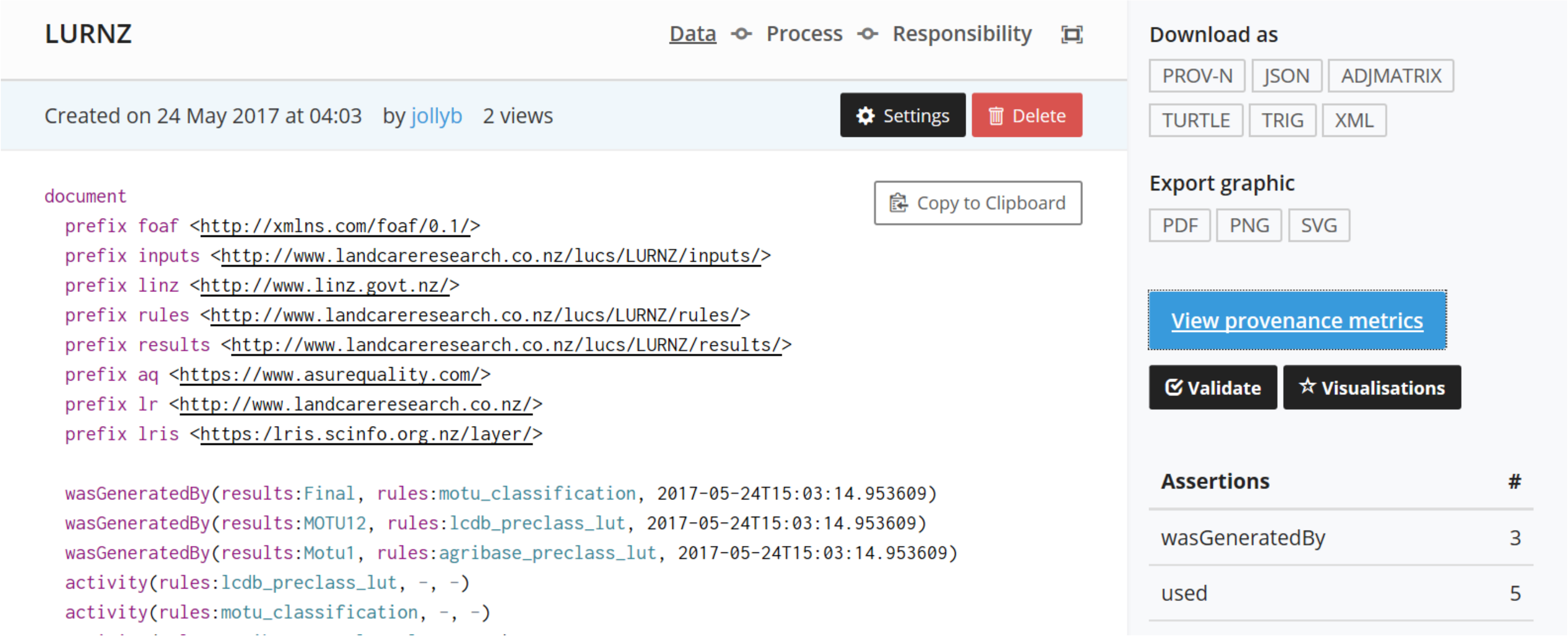
1 Organisations

Short Name	Full Name	URL	LUC Owner
lr	Landcare Research	`http://www.landcareresearch.co.nz/`	True
linz	Land Information New Zealand	`http://www.linz.govt.nz/`	False
aq	AsureQuality	`https://www.asurequality.com/`	False

2 People

Short Name	Full Name	Affiliation	LUC Author	LUC Operator	Delegators
mandersona	Andrew Manderson	Landcare Research	True	False	
ben	Ben Jolly	Landcare Research	False	True	Andrew Manderson
mullerm	Markus Muller	Landcare Research	False	False	Andrew Manderson

Provenance (ProvStore)



LURNZ

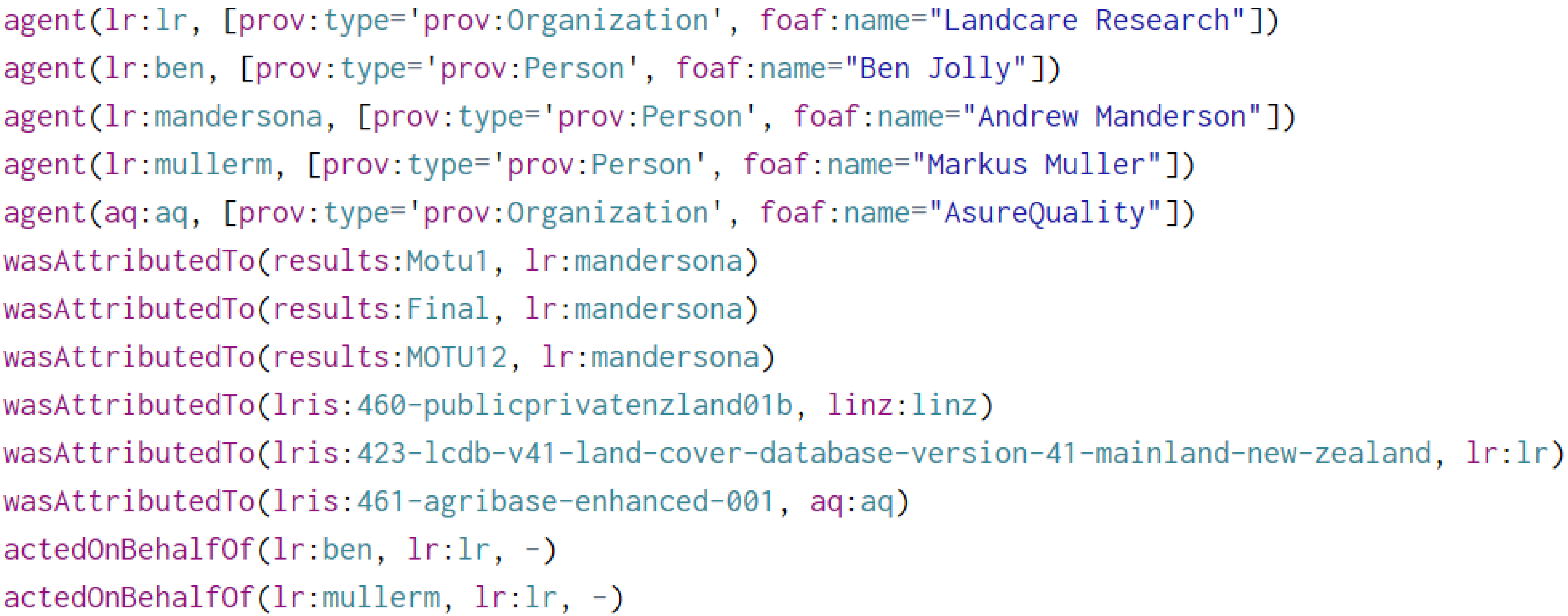
Created on 24 May 2017 at 04:03 by jollyb 2 views

```
document
  prefix foaf <http://xmlns.com/foaf/0.1/>
  prefix inputs <http://www.landcareresearch.co.nz/lucs/LURNZ/inputs/>
  prefix linz <http://www.linz.govt.nz/>
  prefix rules <http://www.landcareresearch.co.nz/lucs/LURNZ/rules/>
  prefix results <http://www.landcareresearch.co.nz/lucs/LURNZ/results/>
  prefix aq <https://www.asurequality.com/>
  prefix lr <http://www.landcareresearch.co.nz/>
  prefix lris <https:/lris.scinfo.org.nz/layer/>

  wasGeneratedBy(results:Final, rules:motu_classification, 2017-05-24T15:03:14.953609)
  wasGeneratedBy(results:MOTU12, rules:lcdb_preclass_lut, 2017-05-24T15:03:14.953609)
  wasGeneratedBy(results:Motu1, rules:agribase_preclass_lut, 2017-05-24T15:03:14.953609)
  activity(rules:lcdb_preclass_lut, -, -)
  activity(rules:motu_classification, -, -)
```

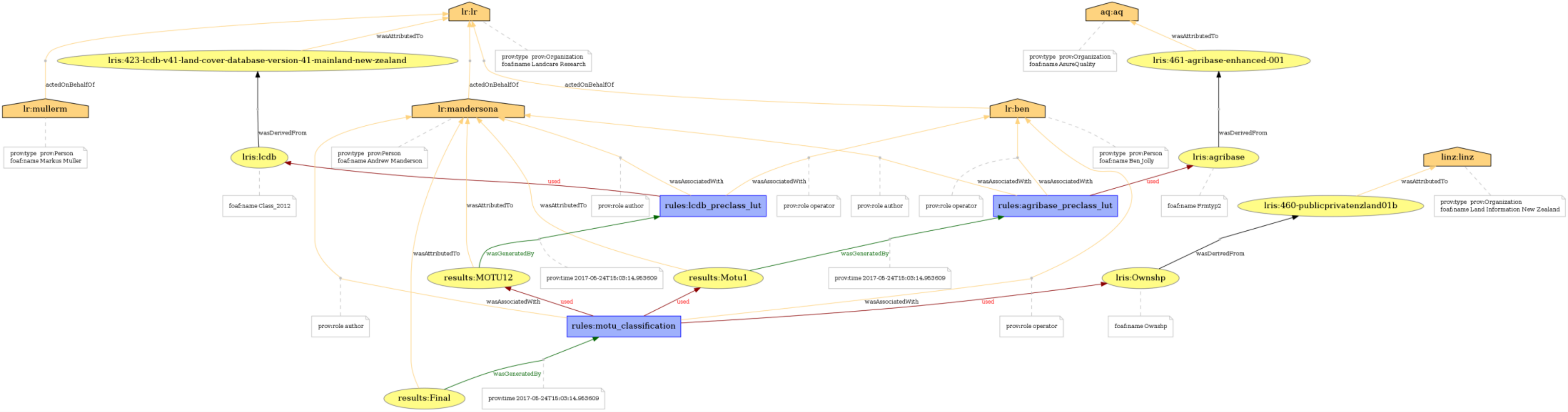
Assertions	#
wasGeneratedBy	3
used	5

Provenance (ProvStore)

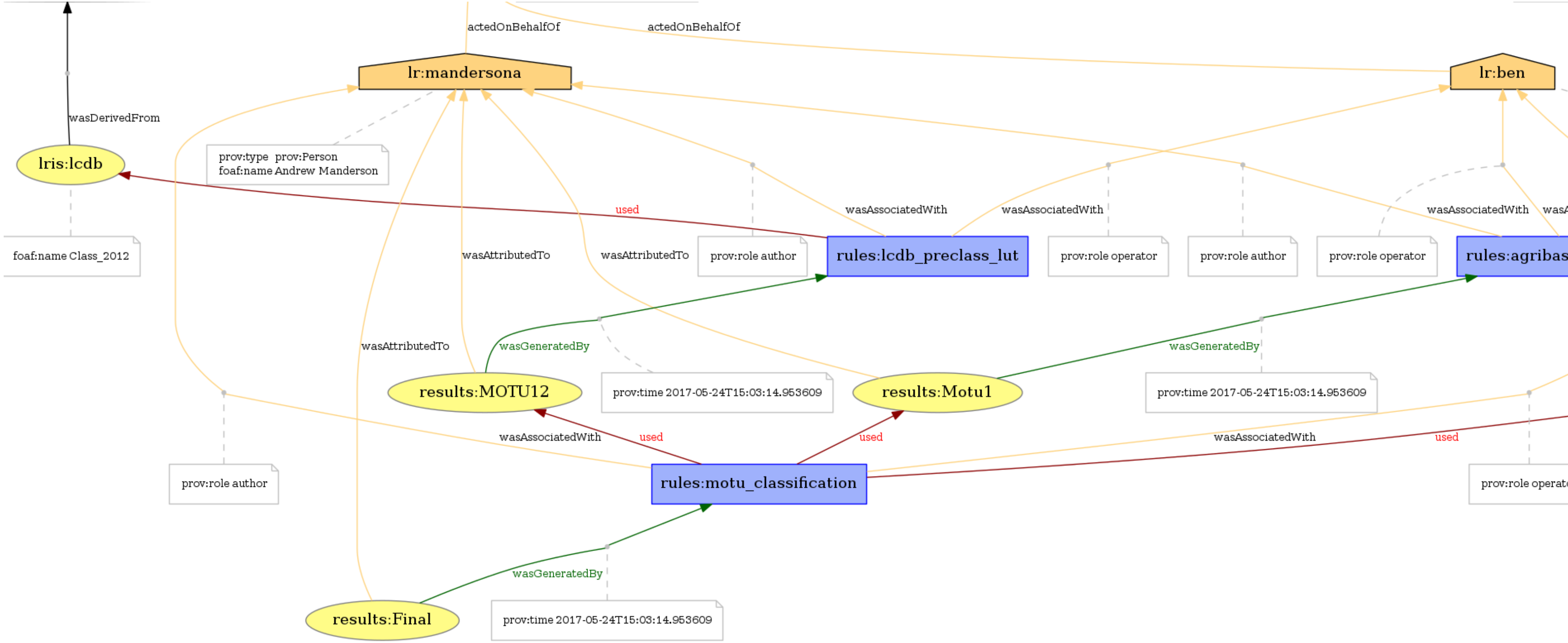


```
agent(lr:lr, [prov:type='prov:Organization', foaf:name="Landcare Research"])
agent(lr:ben, [prov:type='prov:Person', foaf:name="Ben Jolly"])
agent(lr:mandersona, [prov:type='prov:Person', foaf:name="Andrew Manderson"])
agent(lr:mullerm, [prov:type='prov:Person', foaf:name="Markus Muller"])
agent(aq:aq, [prov:type='prov:Organization', foaf:name="AsureQuality"])
wasAttributedTo(results:Motu1, lr:mandersona)
wasAttributedTo(results:Final, lr:mandersona)
wasAttributedTo(results:MOTU12, lr:mandersona)
wasAttributedTo(lris:460-publicprivatenzland01b, linz:linz)
wasAttributedTo(lris:423-lcdb-v41-land-cover-database-version-41-mainland-new-zealand, lr:lr)
wasAttributedTo(lris:461-agribase-enhanced-001, aq:aq)
actedOnBehalfOf(lr:ben, lr:lr, -)
actedOnBehalfOf(lr:mullerm, lr:lr, -)
```

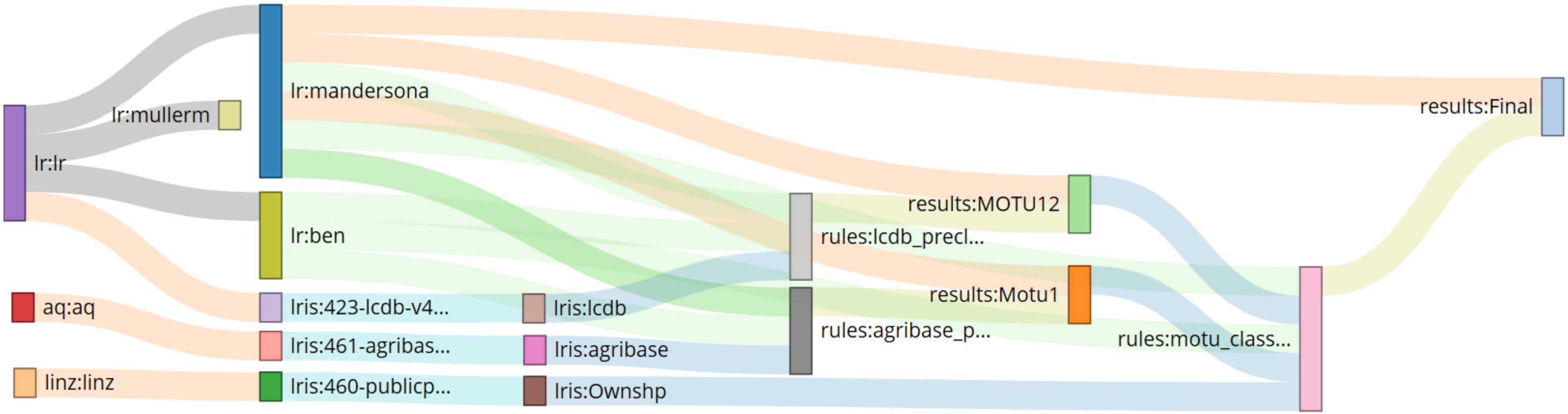
Provenance (ProvStore)



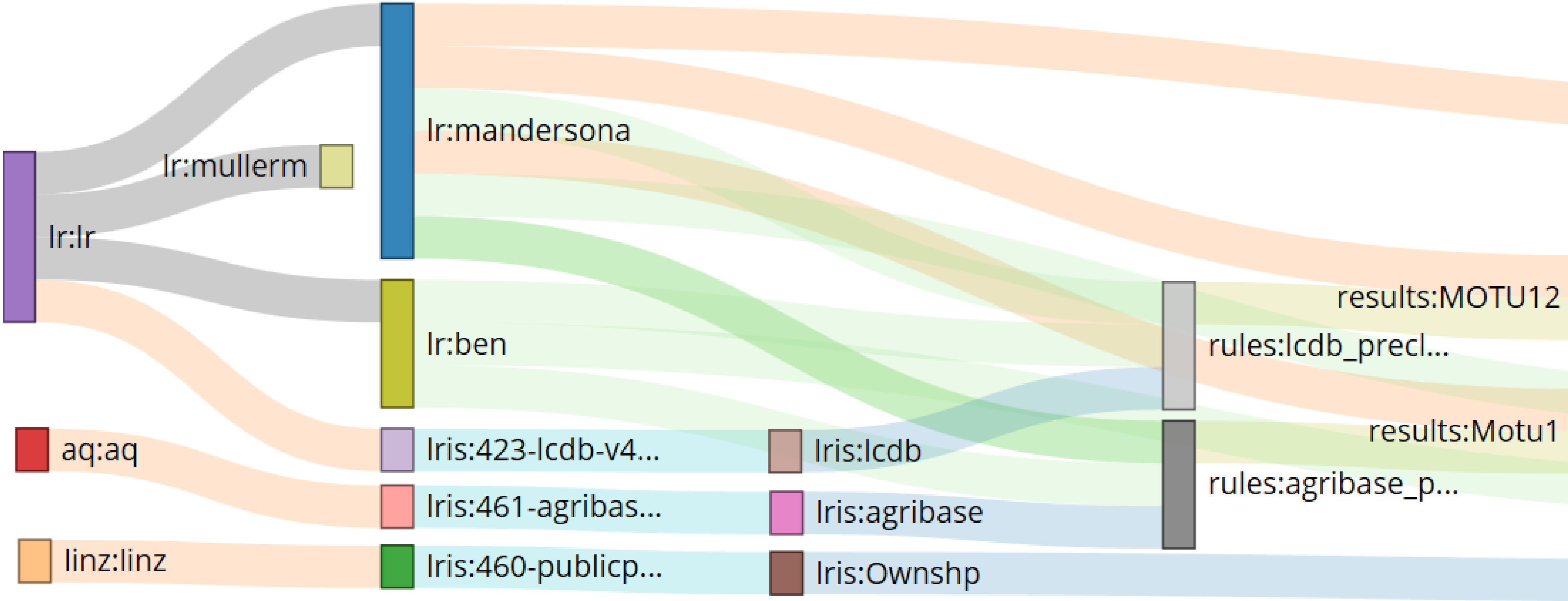
Provenance (ProvStore)



Provenance (ProvStore)



Provenance (ProvStore)



Where is pyluc going?

- Beyond LUCs
- GUI development to make script creation easier
- Beyond LRIS (when Koordinates are ready)
- Beyond Koordinates if the need is there