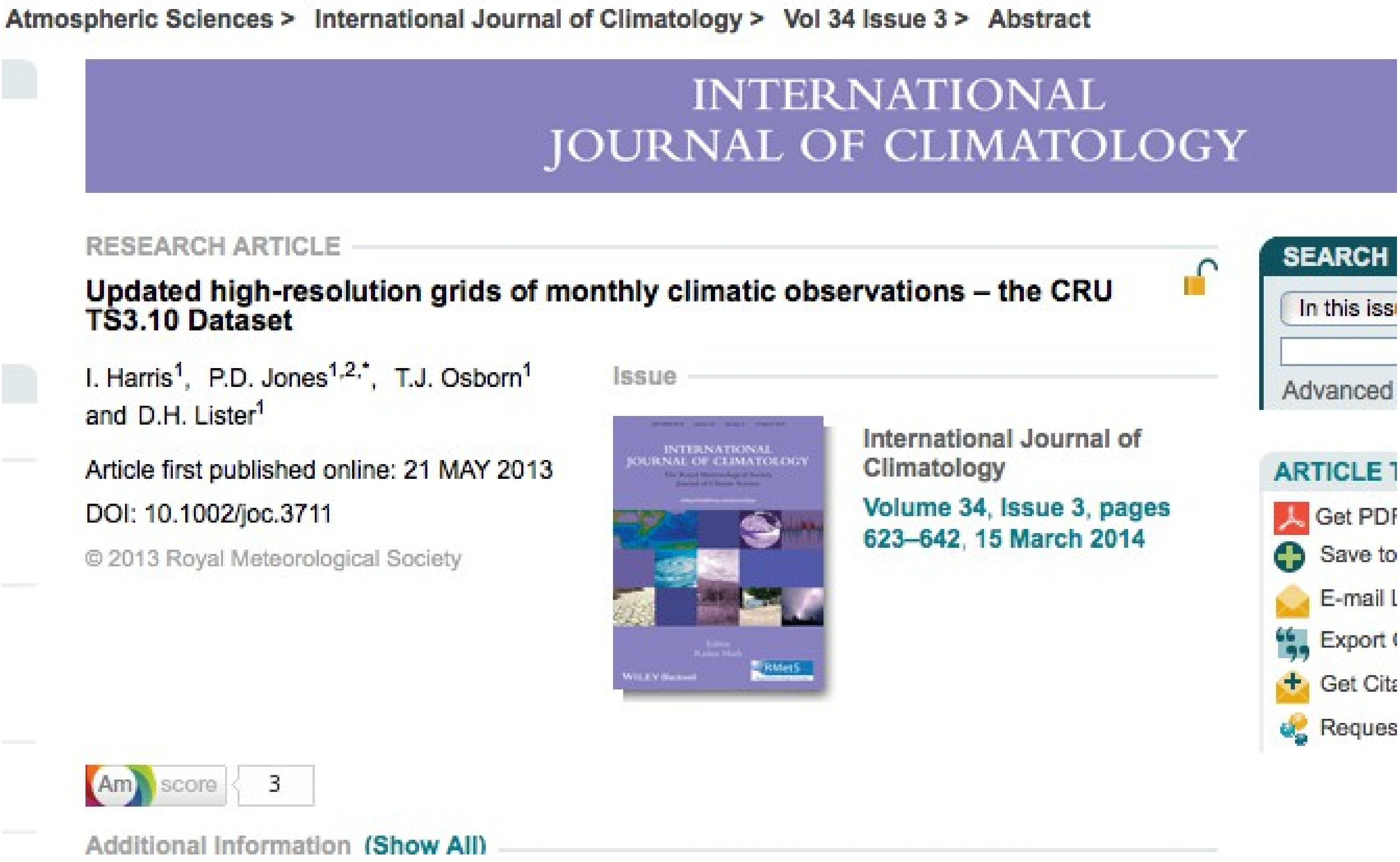
Goal

Capture provenance data relating to the “climategate” data using the W3C provenance model with Annalist

Source



Atmospheric Sciences > International Journal of Climatology > Vol 34 Issue 3 > Abstract

INTERNATIONAL JOURNAL OF CLIMATOLOGY

RESEARCH ARTICLE

Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset

I. Harris[1], P.D. Jones[1,2,*], T.J. Osborn[1] and D.H. Lister[1]

Article first published online: 21 MAY 2013

DOI: 10.1002/joc.3711

Issue

International Journal of Climatology

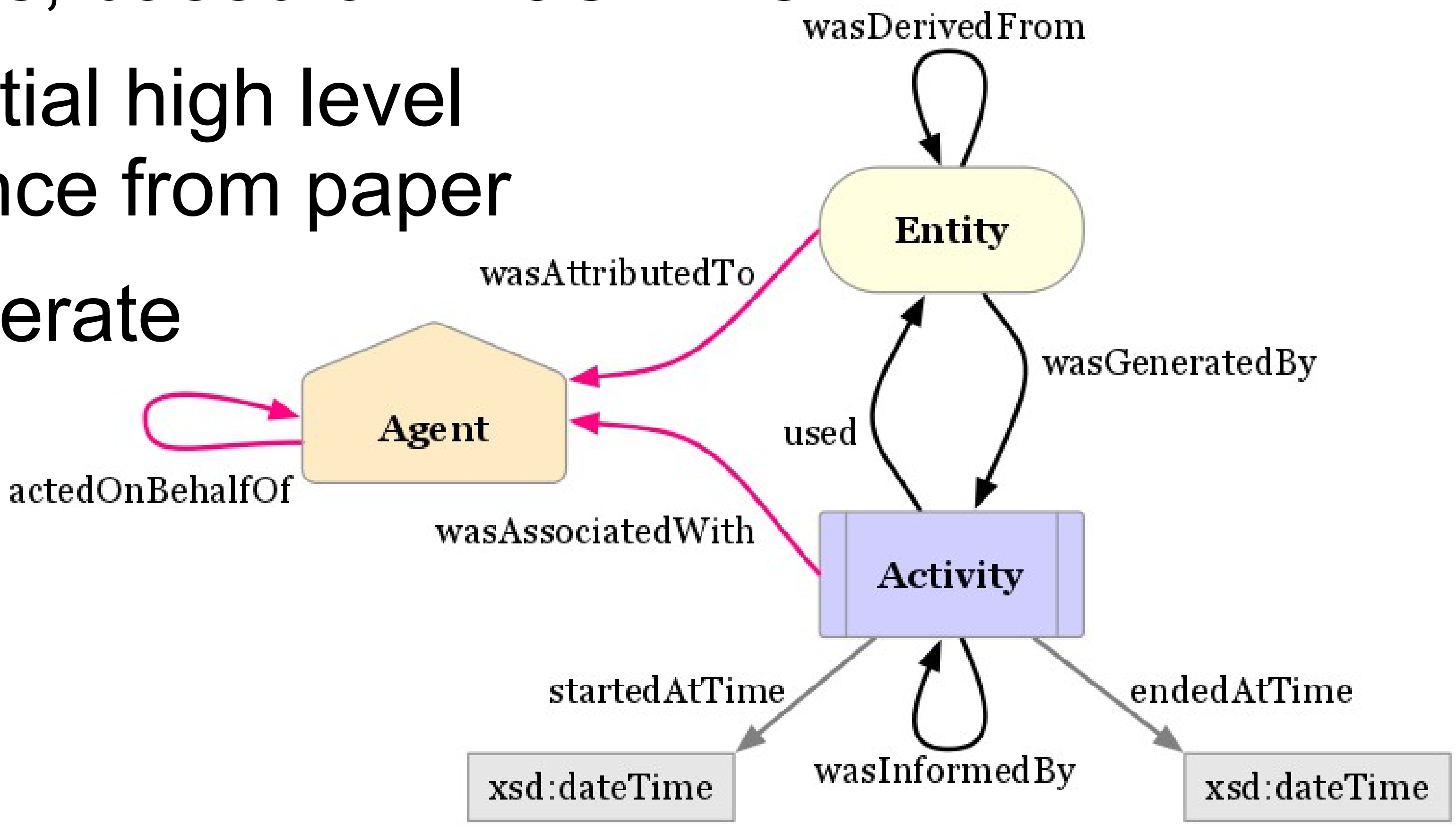
Volume 34, Issue 3, pages 623–642, 15 March 2014

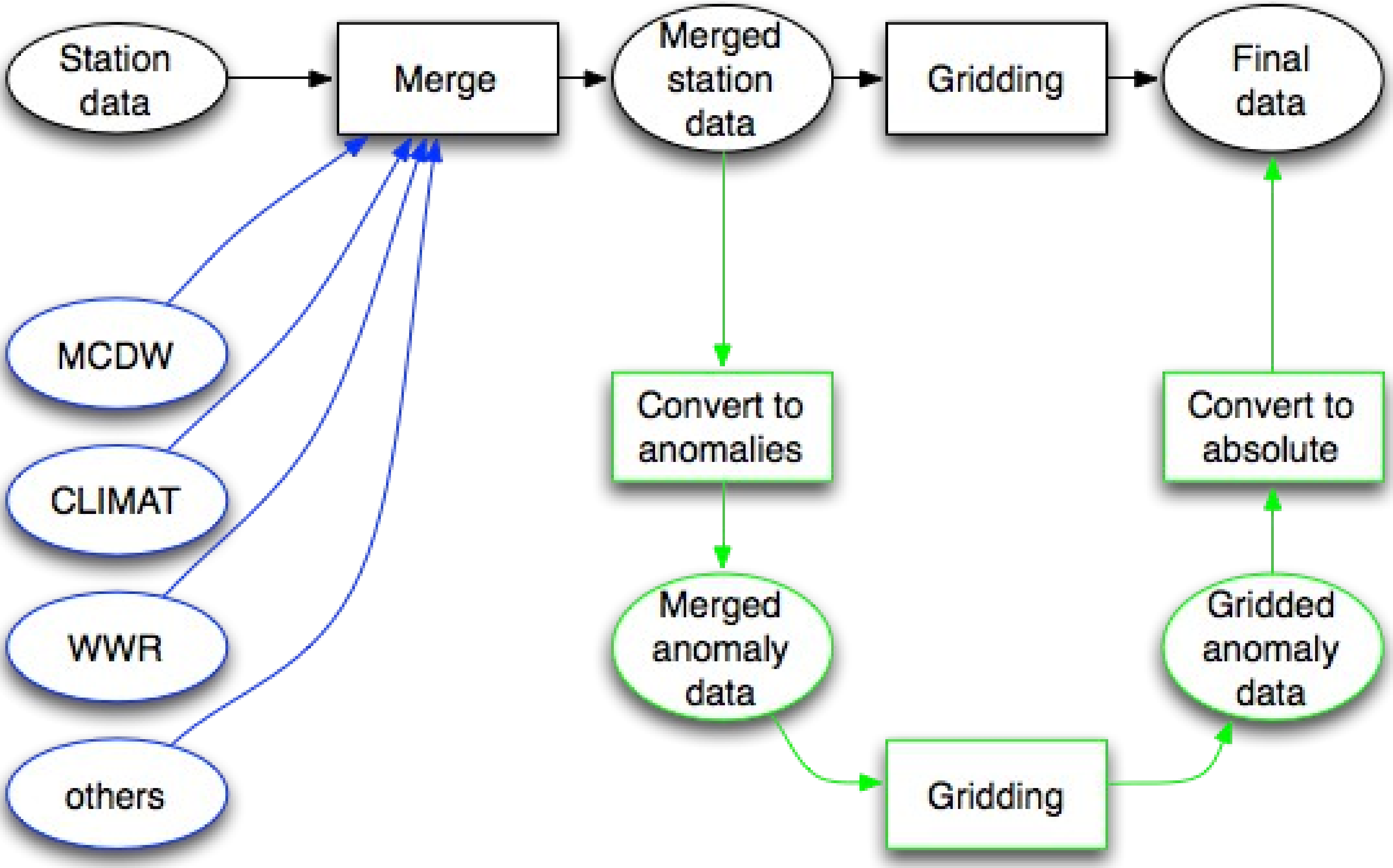
Additional Information (Show All)

http://dx.doi.org/10.1002/joc.3711

What we did

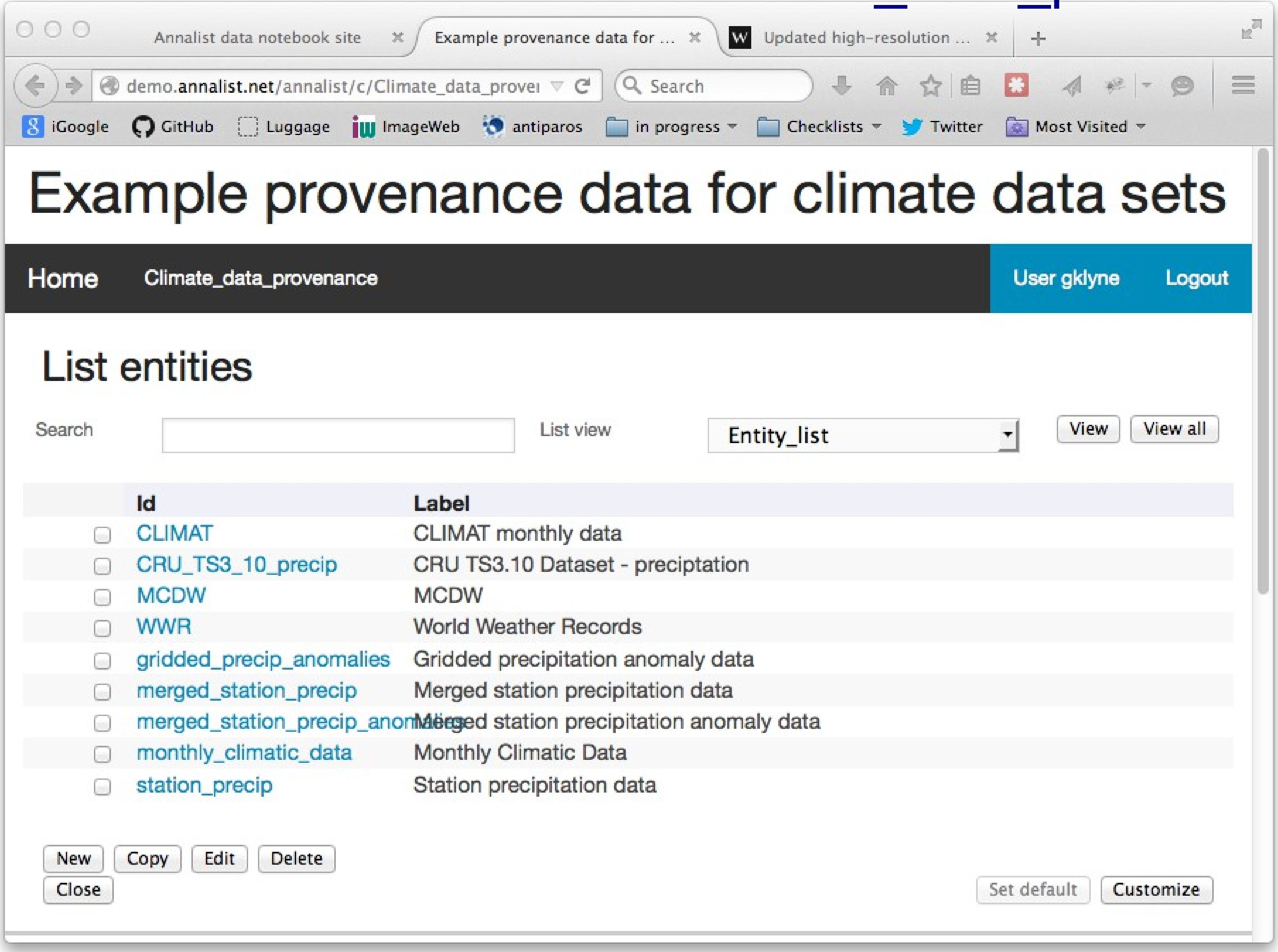
1. Create Annalist collection and GitHub repo
2. Created initial Annalist data definitions, based on W3C PROV
3. Glean initial high level provenance from paper
4. Refine, iterate




Station data
Merge
Merged station data
Gridding
Final data
MCDW
CLIMAT
WWR
others
Convert to anomalies
Merged anomaly data
Gridding
Gridded anomaly data
Convert to absolute

Outcome 1 - online provenance capture/demo tool

http://demo.annalist.net/annalist/c/Climate_data_provenance/



Example provenance data for climate data sets

List entities

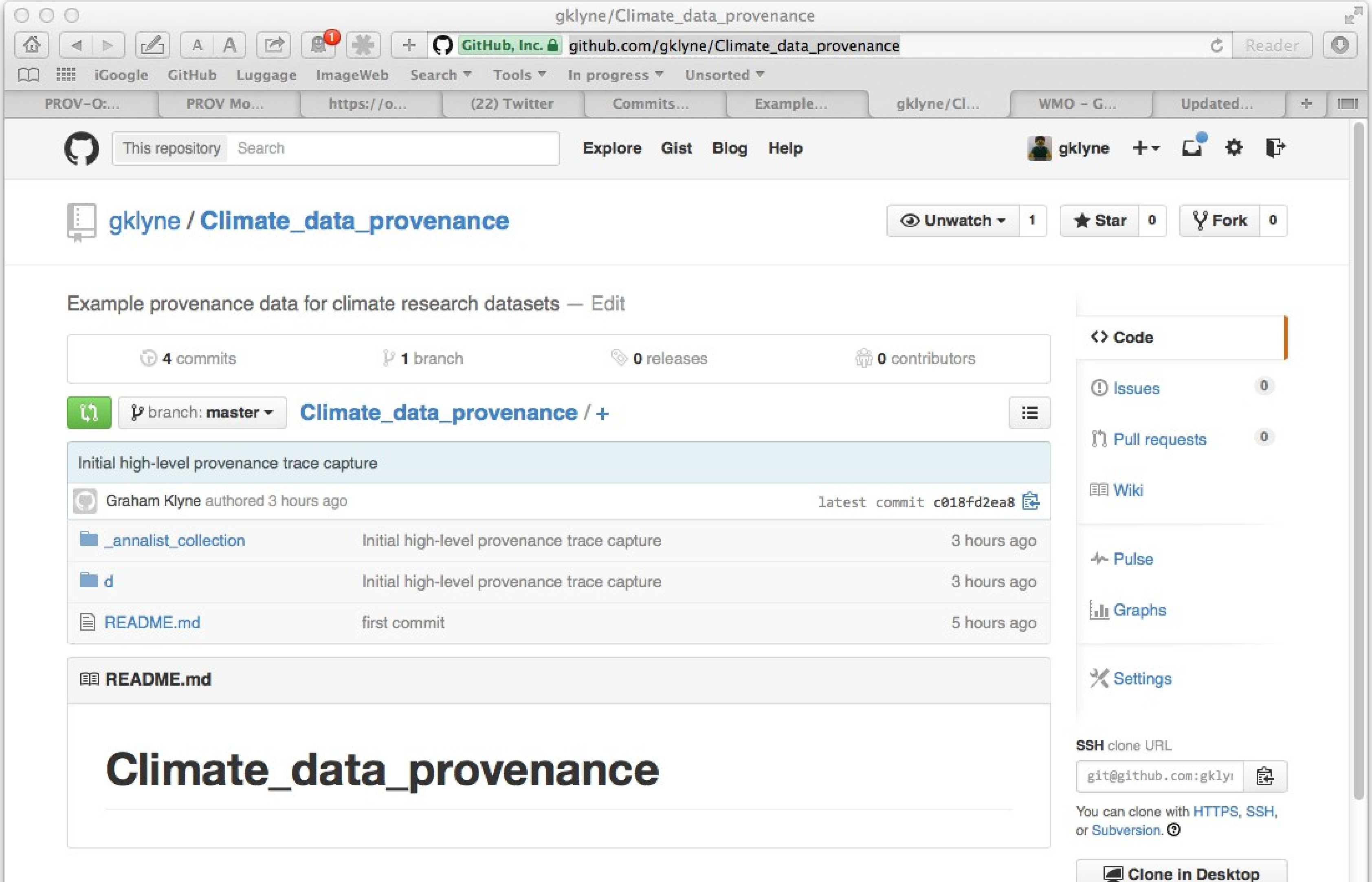
Search

List view: Entity_list

	Id	Label
□	CLIMAT	CLIMAT monthly data
□	CRU_TS3_10_precip	CRU TS3.10 Dataset - preciptation
□	MCDW	MCDW
□	WWR	World Weather Records
□	gridded_precip_anomalies	Gridded precipitation anomaly data
□	merged_station_precip	Merged station precipitation data
□	merged_station_precip_anom	Merged station precipitation anomaly data
□	monthly_climatic_data	Monthly Climatic Data
□	station_precip	Station precipitation data

Outcome 2 – Annalist data definitions (JSON) in GitHub

https://github.com/gklyne/Climate_data_provenance



gklyne / **Climate_data_provenance**

Example provenance data for climate research datasets — Edit

4 commits **1** branch **0** releases **0** contributors

branch: **master** ▾ **Climate_data_provenance** / +

Initial high-level provenance trace capture

Graham Klyne authored 3 hours ago `latest commit c018fd2ea8`

_annalist_collection	Initial high-level provenance trace capture	3 hours ago
d	Initial high-level provenance trace capture	3 hours ago
README.md	first commit	5 hours ago

README.md

Climate_data_provenance

SSH clone URL

`git@github.com:gklyı`

You can clone with HTTPS, SSH, or Subversion.

Summary

- Manual provenance extraction is tedious
  - automated capture would be much easier!
- Have created initial set of Annalist definitions for provenance
  - next time will be so much easier
- It proved very effective to start with a very high level view of provenance, and iterate
  - be agile
- Learned more about areas where Annalist usability still needs work