UNIVERSITY OF CAMBRIDGE

OSCAR Tagging of Atmospheric Science Papers

A Quick Overview

Hannah Barjat, Lezan Hawizy,
Peter Murray-Rust

Aims

- Increase visibility of atmospheric science to the general public.
- Increase speed with which literature searches can be carried out.
- Provide easier access to published field data for modelers.
- Show connectivity of research campaigns – could extend outside of atmospheric research.

Aims

- Specific example
  - Demonstrate record of ACP publications arising from studies at field stations.
  - Demonstrate affiliations of authors using field stations and publishing in ACP.
  - Demonstrate publication of campaign data in ACP.

Method: Choice of Publication

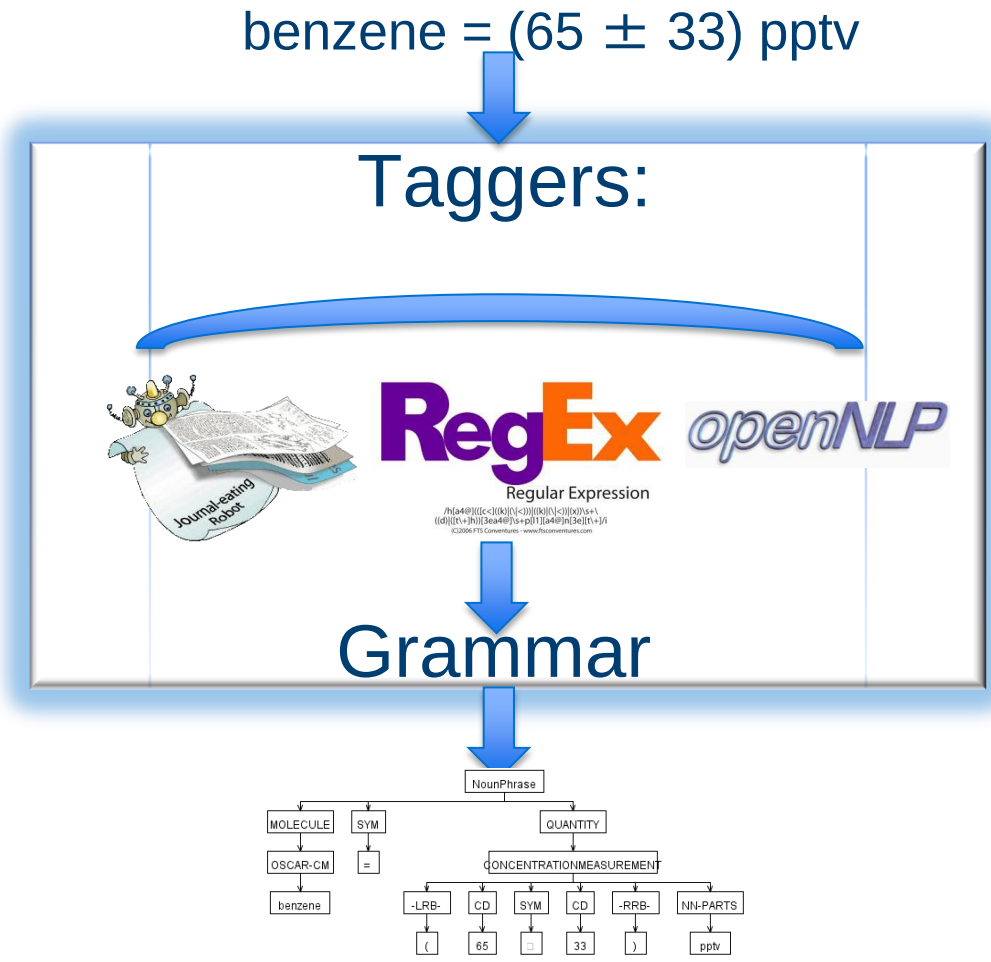
- Atmospheric Chemistry and Physics a publication of the European Geophysical Union.
- Publication freely available; including an XML version of the Abstracts.
- Range of papers: field, lab and modelling studies.

Method: Choice of Example Abstracts

- ACP searched for “ACP Abstract” and “station”
  - An example where we’d have a good chance of mapping data.
- XML versions of Abstracts downloaded – thanks to Sam Adams
  - Further refinement to eliminate Abstracts that had “station” only within the affiliations or references.
  - 200 Abstracts

Method: ChemicalTagger

- OpenSource NLP tool for processing chemical text.
- Combines Chemical Entity Recognition (OSCAR) with NLP techniques
- Extendible and Reconfigurable Taggers and Parsers



Method: Adapting ChemicalTagger

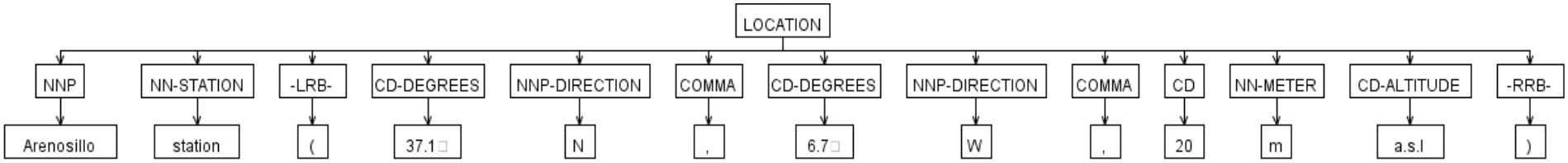
- Dictionaries
  - IUPAC
  - Met Glossary
  - Station Co-ordinates
  - ACRONYMS
  - Units – related work on unit dictionaries (Joe Townsend)

Method: Adapting ChemicalTagger

- Constructed Atmospheric Chemistry Grammar to extract :
  - Location
  - Time
  - Parenthetical phrases
  - Acronyms
  - Campaigns
  - Action Phrases: Measurement, Observation
  - Mixing ratio (to be extended to cover a wider range of units).

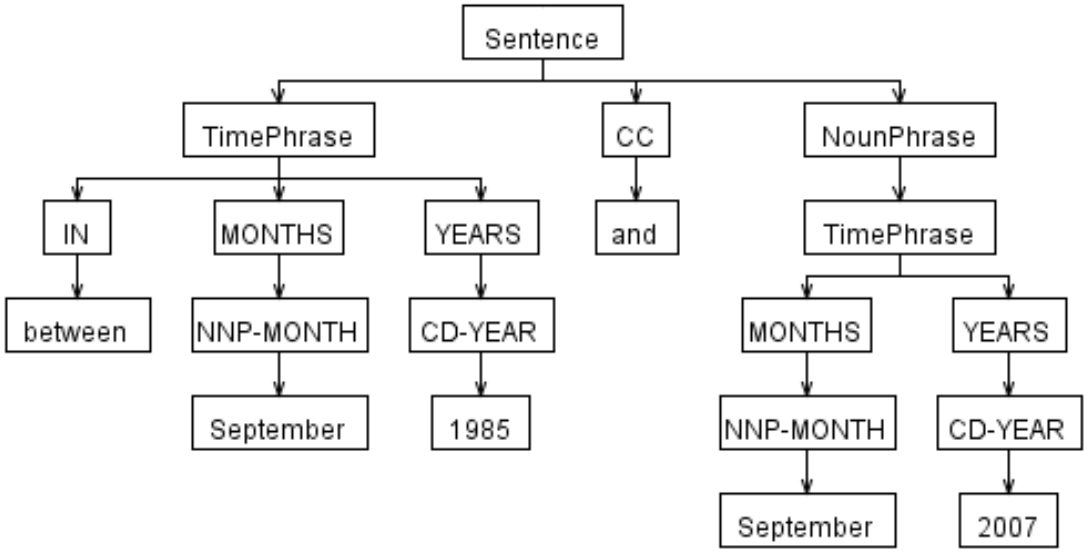
Example Results

- Location: station, country, degrees (latitude, longitude), altitude units (m a.s.l etc.)
- *E.g: Arenosillo station (37.1° N, 6.7° W, 20 m a.s.l)*



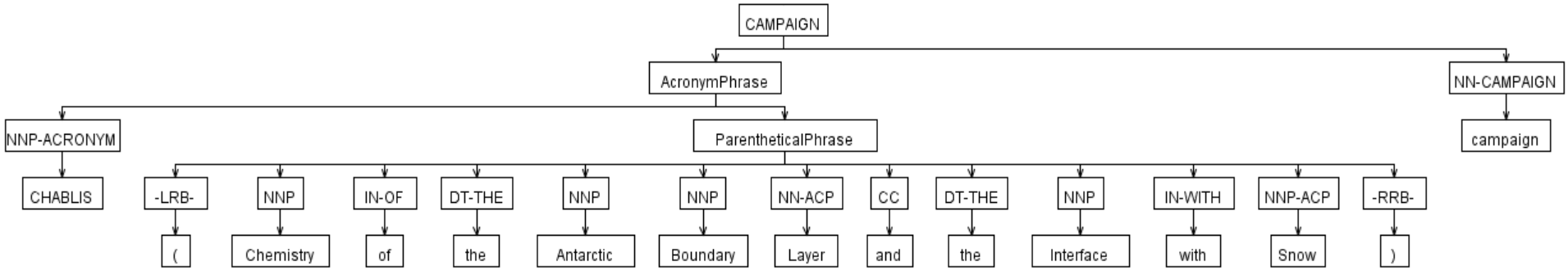
Example Results

- Time: time phrases, year, month, day.
- *E.g. between September 1985 and September 2007.*



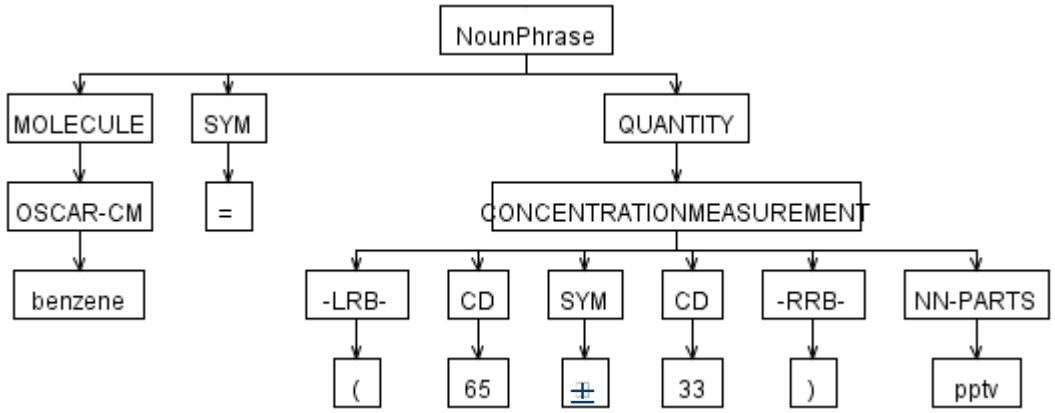
Example Results

- Campaign Names:
- E.g. ……CHABLIS (Chemistry of the Antarctic Boundary Layer and the Interface with Snow) campaign ….



Example Results

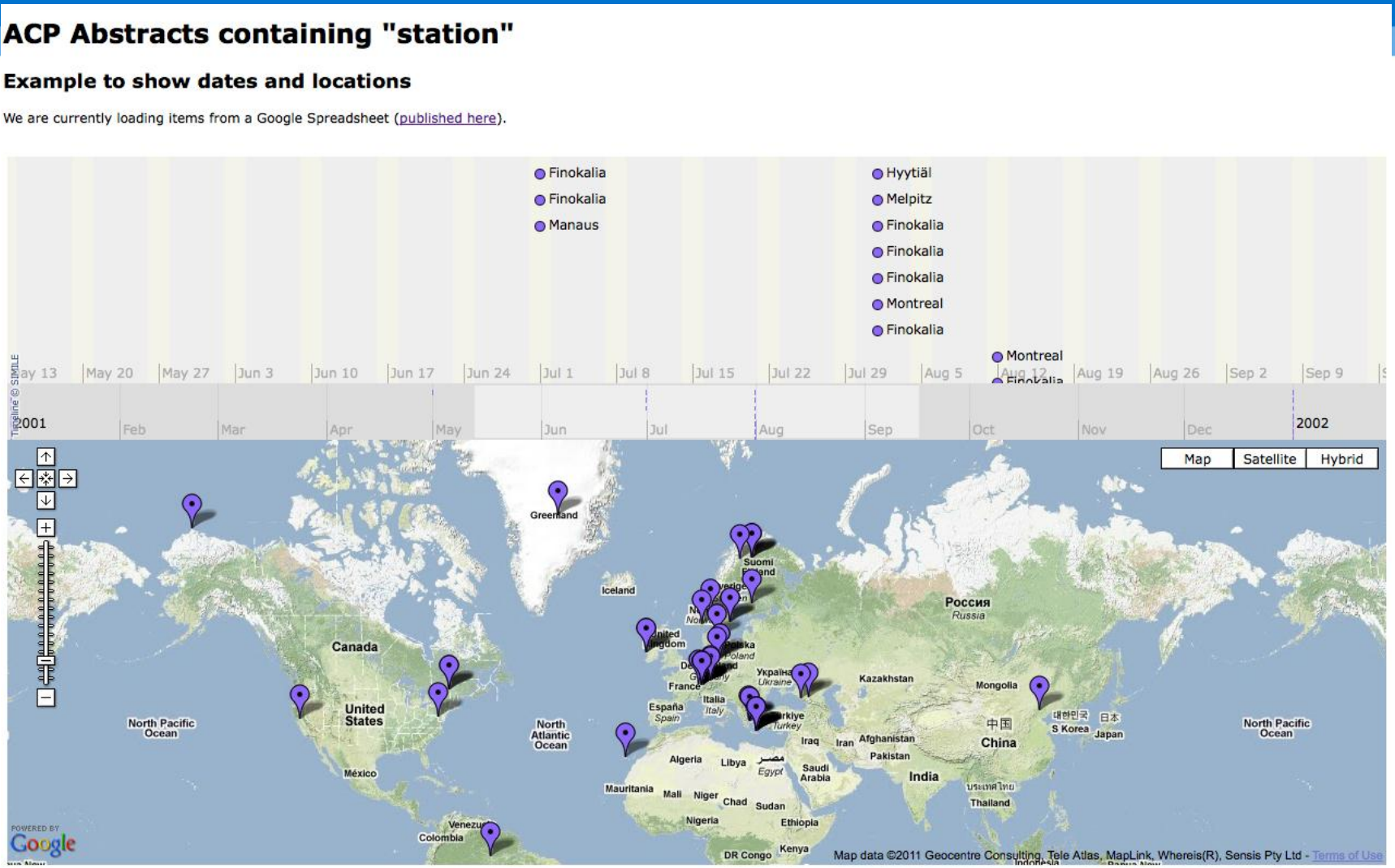
- Action Phrases: Measurement, Observation
  - Gives confidence that molecules are being measured i.e. can distinguish from model results.
- Mixing ratio: benzene = (65 ± 33) pptv



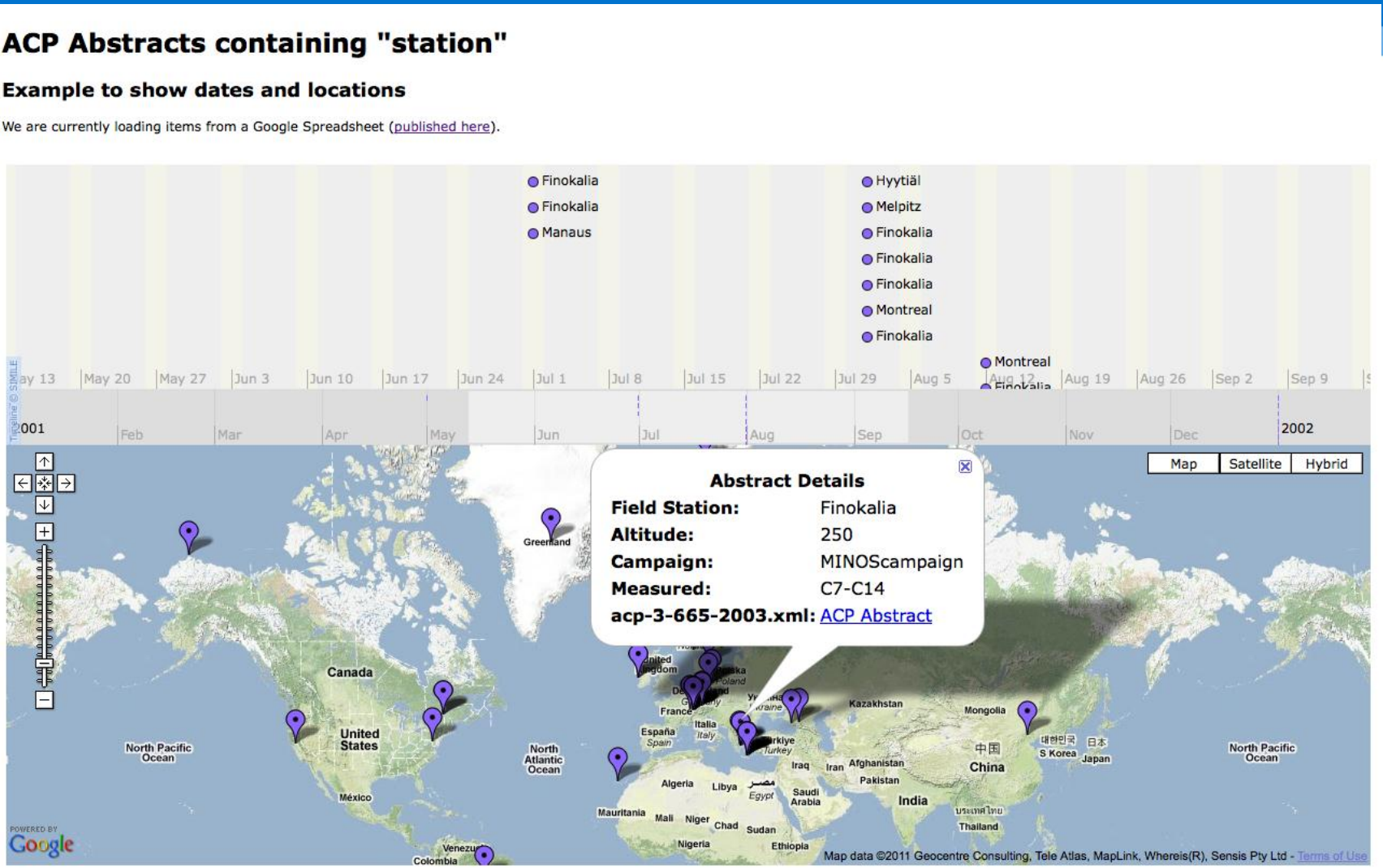
Visualisation

- Using Google Spreadsheet and Google Maps.
  - Straightforward provided data is in correct format for a spreadsheet.
  - Could improve with a little work to extract data differently.

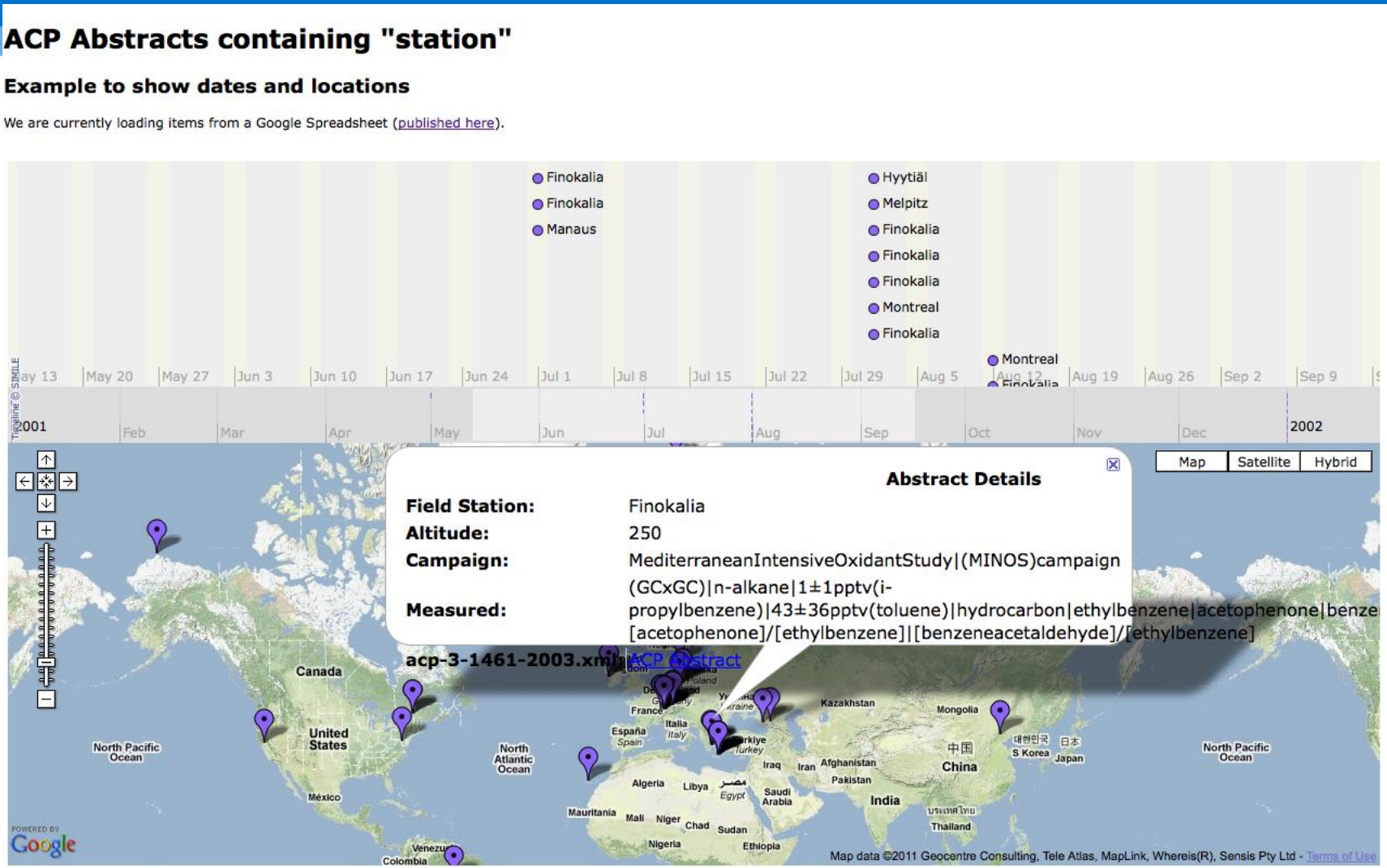
Visualisation



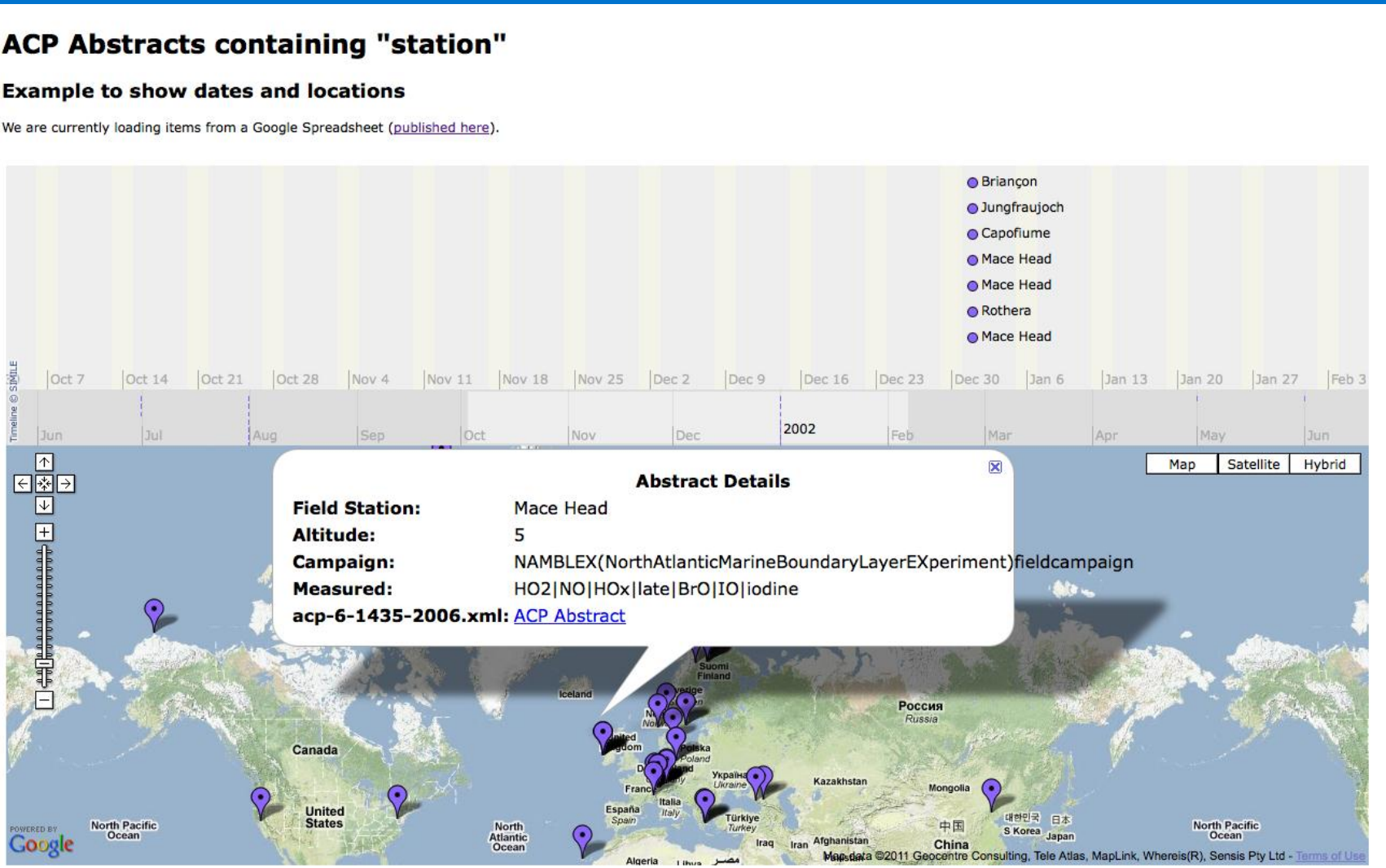
Visualisation



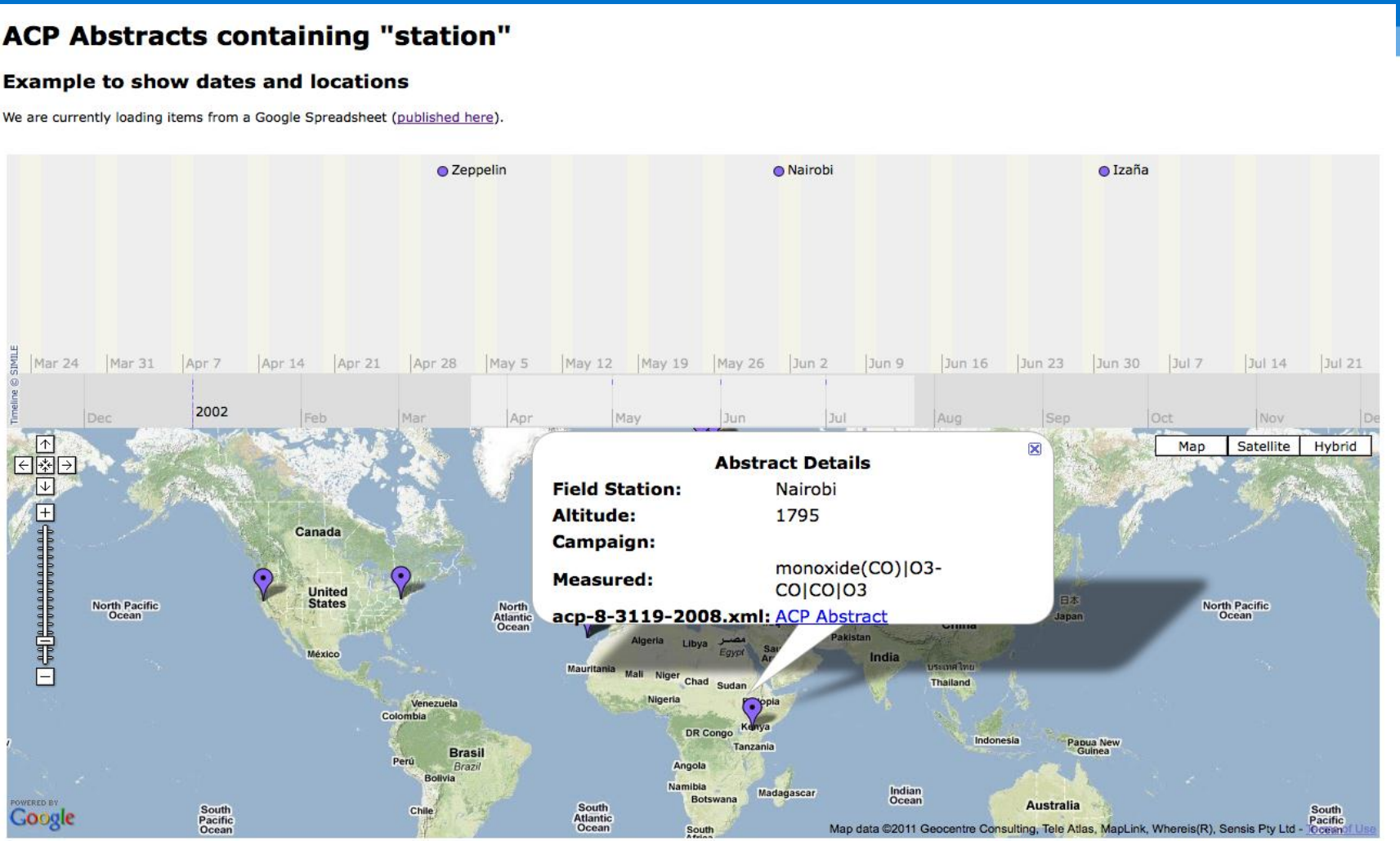
Visualisation



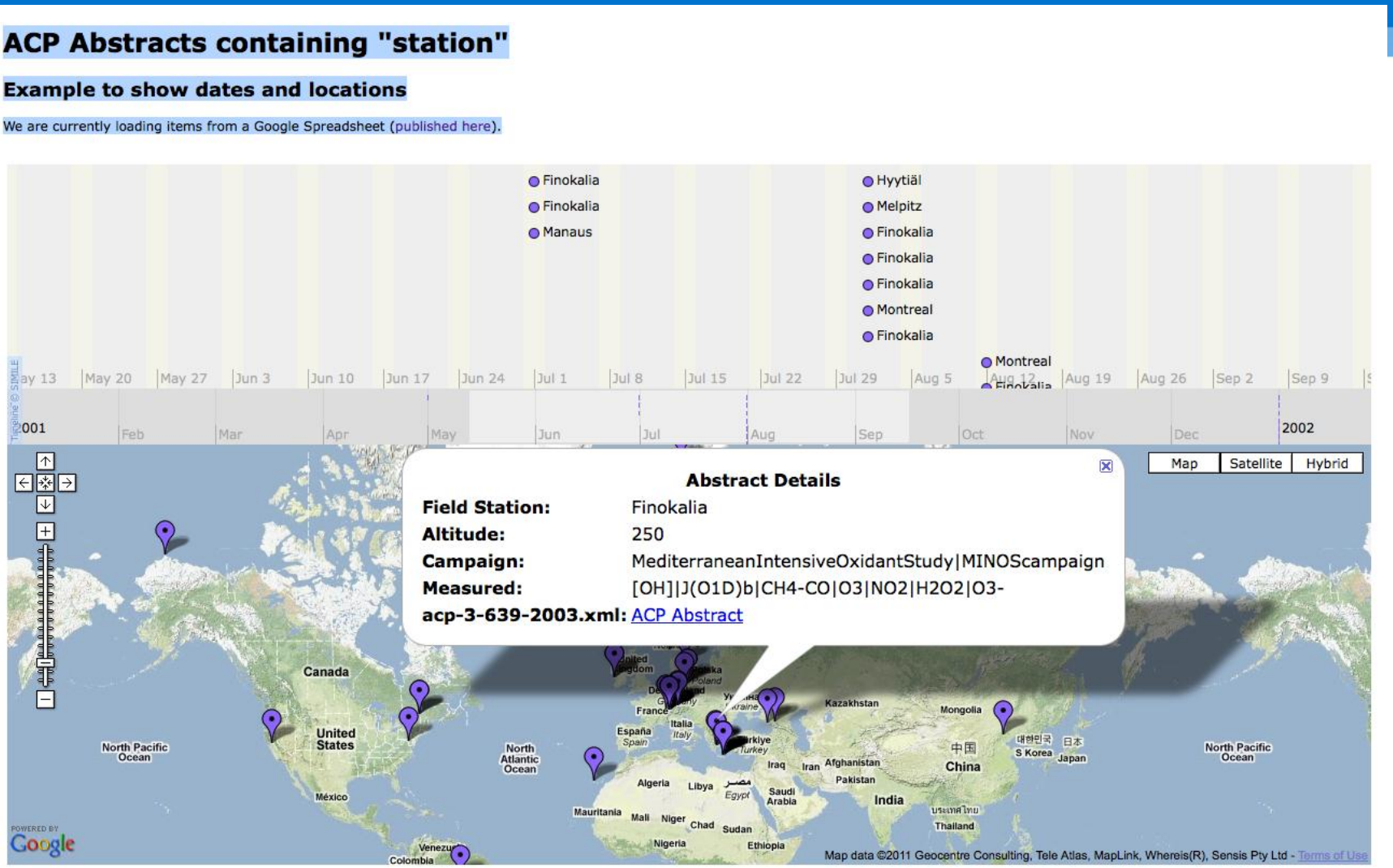
Visualisation



Visualisation



Visualisation



Results

- For examples chosen we found:
- 60% of Abstracts had a clearly identifiable station location. Only 1 false positive found (Montreal – as in the Protocol).
- 70% of Abstracts had a clearly identifiable date(s). Few false positives.
- Measured species – a number of false positives, as well as correct matching to molecules.
- We are able to pick out mixing ratios.
  - Further work to see how useful this is depends on associated words e.g. “background’, maxima, boundary layer etc.

Conclusions

- We have the basis of a tool for quickly viewing ACP Abstracts:
  - To pick out:
  - Location
  - Dates
  - Field station and campaign names
  - Measurements
  - Further work needs to be carried out to (1) improve data extraction and reliability, (2) improve testing and (3) improve visualisation.

Acknowledgements

- The Peter Murray Research group
  - In particular, Sam Adams and Joe Townsend.
  - EPSRC
  - ACP Editors (Rolf Sander and Ulrich Pöschl) for their encouragement.