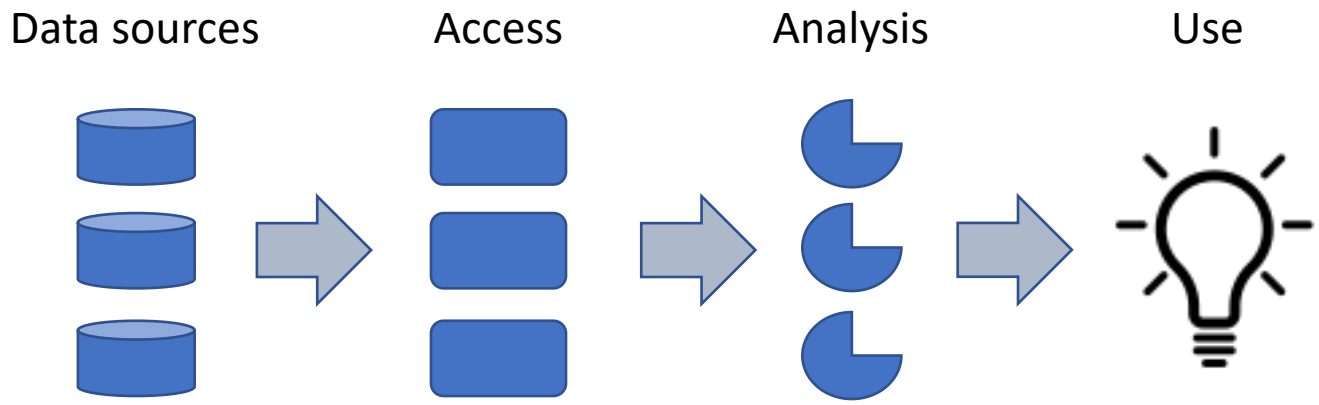
Session 1: Tools, data, methods

Chair: Jari Arkko

Presentations

- Datatracker interface (Sparks)
- BigBang (Benthall)
- SODESTREAM (McQuistin)
- IETF website analytics (Wood)



Relevant papers:

- Using Complex Systems Analysis to Identify Organizational Interventions (Sebastian Benthall)
- The ietfdata Library (Stephen McQuistin, Colin Perkins)
- The RFC Prolog Database (Marc Petit-Huguenin)
- Observations about IETF process measurements (Jari Arkko)
- And this, though not a paper: https://www.ietf.org/policies/web-analytics/ (IETF)

Accessing Datatracker Data

Robert Sparks

IAB AID Workshop

Session 1

What's available?

- Files (drafts, RFCs, agendas, minutes, photos)
  - Available over HTTPS and through rsync
- Metadata about People, Documents, Groups, Meetings, and more
  - Stored in SQL, structured as Django Models
- Archives of mailing lists
  - Managed by the mailarchive Django project rather than the datatracker

Details at https://notes.ietf.org/iab-aid-data-resources

Two ways to get to the datatracker data

- Set up a local development environment
  - Using docker (note that the first build will take 40-ish minutes)
  - Developer database dump refreshed daily
  - Django shell allows construction of arbitrary querysets
  - Instructions at https://notes.ietf.org/iab-aid-data-resources
- Use the v1 API
  - Built on Tastiepie
  - Can just be browsed (xml or json output)
  - Best accessed with curl and jq
  - Not as capable as the development environment
  - Some ordering care needed when retrieving a large number of records

Preview of what you can do

In [3]: Document.objects.filter(documentauthor__person__name= "Robert Sparks").count()

Out[3]: 258

In [4]: Counter(Document.objects.filter(documentauthor__person__name= "Robert Sparks").values_list('type_id',flat=True))

Out[4]: Counter({'draft': 67, 'review': 191})

Preview of what you can do

```
% curl "https://datatracker.ietf.org/api/v1/doc/document/?states__slug=lc&format=json" \
    | jq "[.objects[] | .name]"
[
  "draft-eastlake-rfc6931bis-xmlsec-uris",
  "draft-ietf-i2nsf-capability-data-model",
  "draft-ietf-i2nsf-nsf-monitoring-data-model",
  "draft-ietf-httpbis-priority",
  "draft-ietf-acme-dtnnodeid",
  "draft-ietf-httpbis-http2bis",
  "draft-ietf-lamps-samples"
]
```

What's in there?

- A lot – the datatracker is a large application, with complex data relationships
  - Over 100,000 documents
  - Nearly 20,000 people
  - Over 1000 meetings (including interims)
- Minimal PII for People
  - Names and email addresses
  - Sometimes Affiliation and Country
  - No addresses (though those can sometimes be mined from drafts)
  - No other explicitly captured demographics
- See https://notes.ietf.org/iab-aid-datatracker-database-overview

What's in there: History

- Most metadata is saved when a Document, Group, or Person object is modified
- Reasonably complete for recent (10 to 15 years) history
- Poor for older history – data is often incomplete, and is occasionally completely wrong
- The models were designed for tracking the current state of work. Mining the history records can be complicated.
  - It is very hard, for example, to determine when someone stopped being a chair of a given working group.

Backup slides

What’s in there: Code

- Don’t ignore the codebase
- Many utilities exist to make data mining easier
  - Finding the current (or final) IESG ballot state for a document
  - Extracting authors from text Internet-Drafts
  - Finding the chain of all documents that ultimately normatively depend on a given document through dependencies on other documents.

Getting started

- Explore the development environment and the v1 API
- Ask questions:
  - I’m available at email:rjsparks@nostrum.com and on the IETF Slack
  - Consider subscribing to tools-discuss@ietf.org