B2FIND records in neo4j

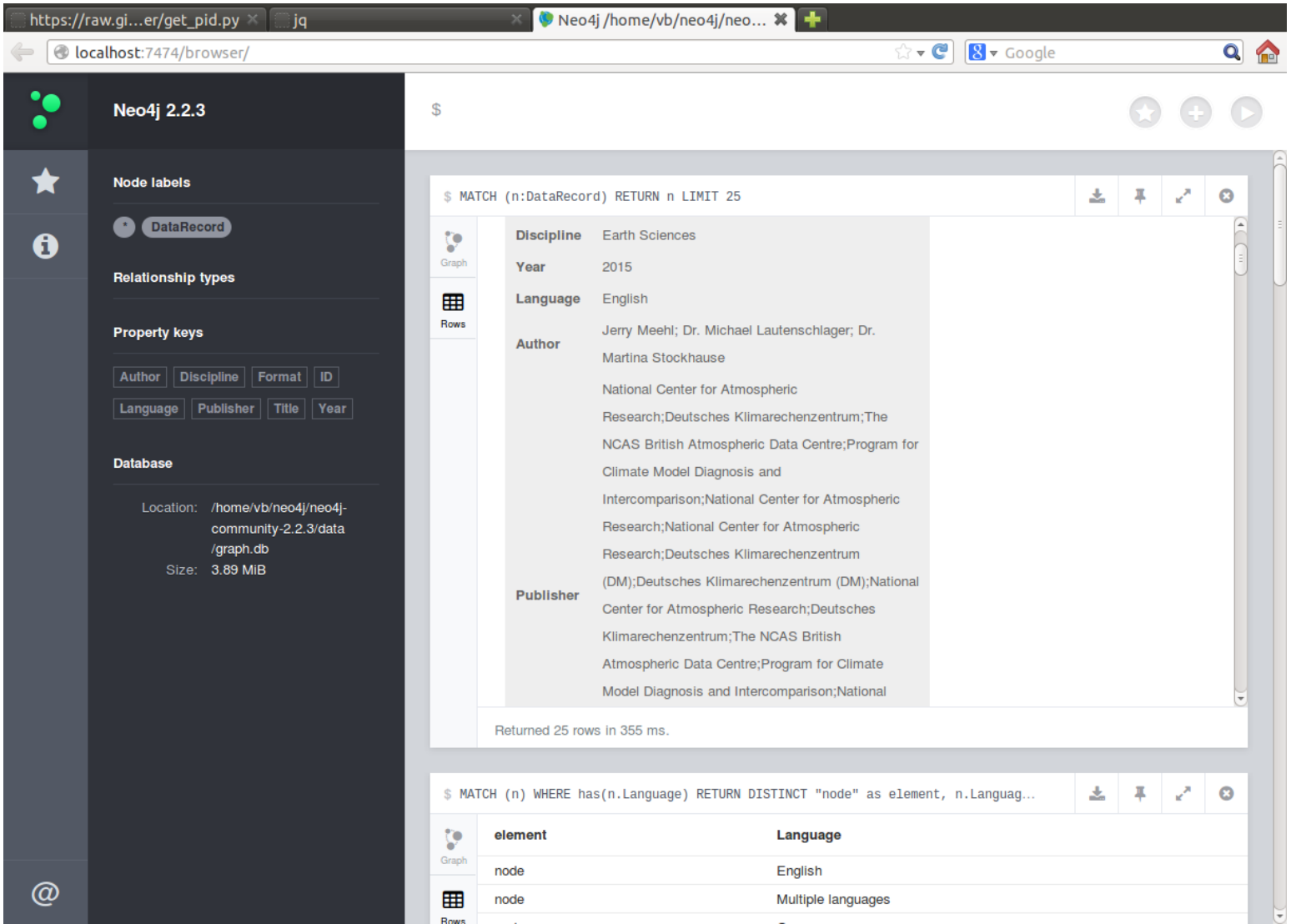
Vasily Bunakov
vasily.bunakov@stfc.ac.uk

EUDAT developers meeting in SURFSara,
Amsterdam, 6-8 October 2015

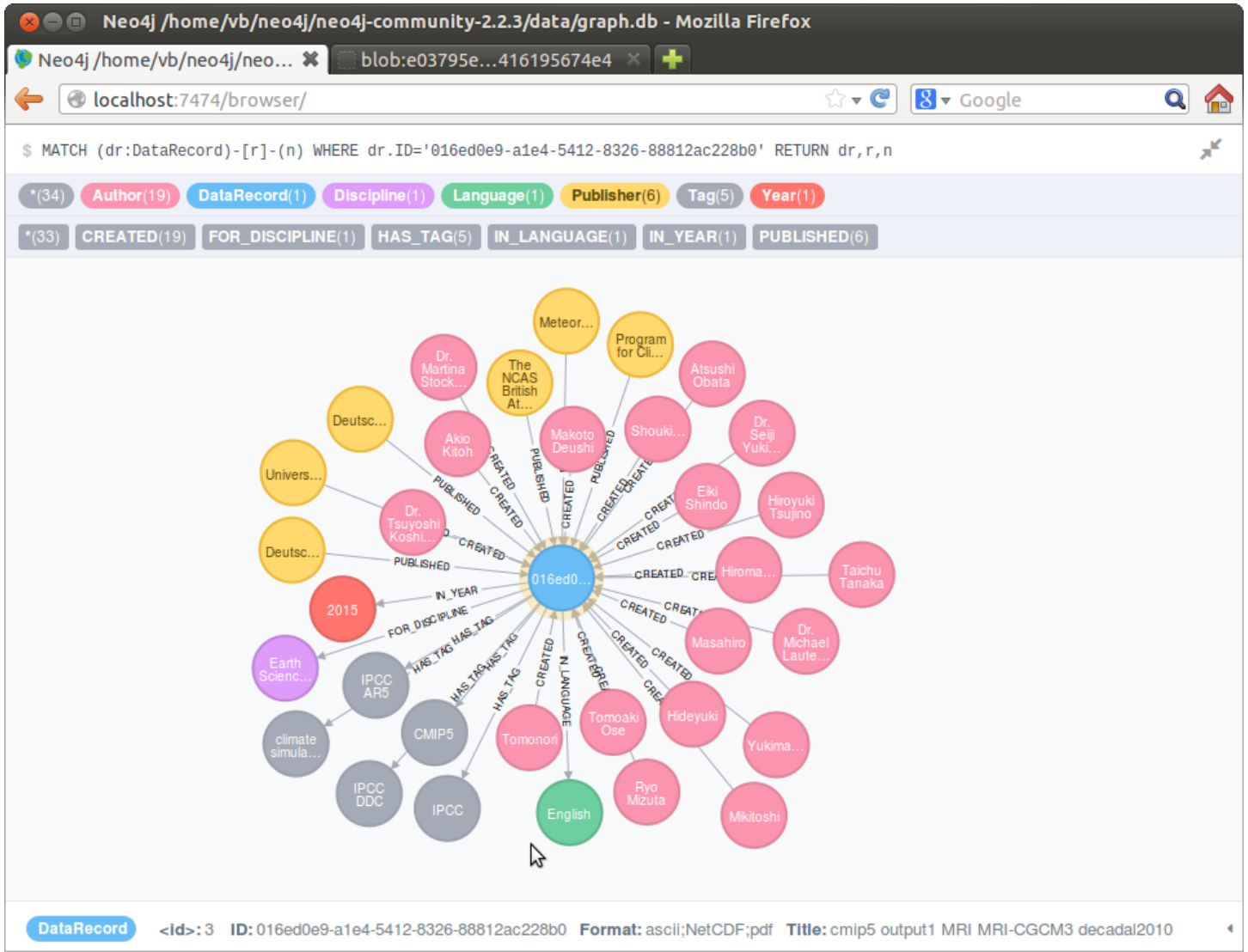
Data migration summary

- ~ 800 B2FIND data records imported in neo4j
- New information entities / graph nodes produced for Authors (491), Publishers(136), Tags(214) and Languages(2)
- **~ 34,000 relationships produced between B2FIND records and newly created entities**
- 10 ORCID records (for a selected Author) imported and mixed up with B2FIND records

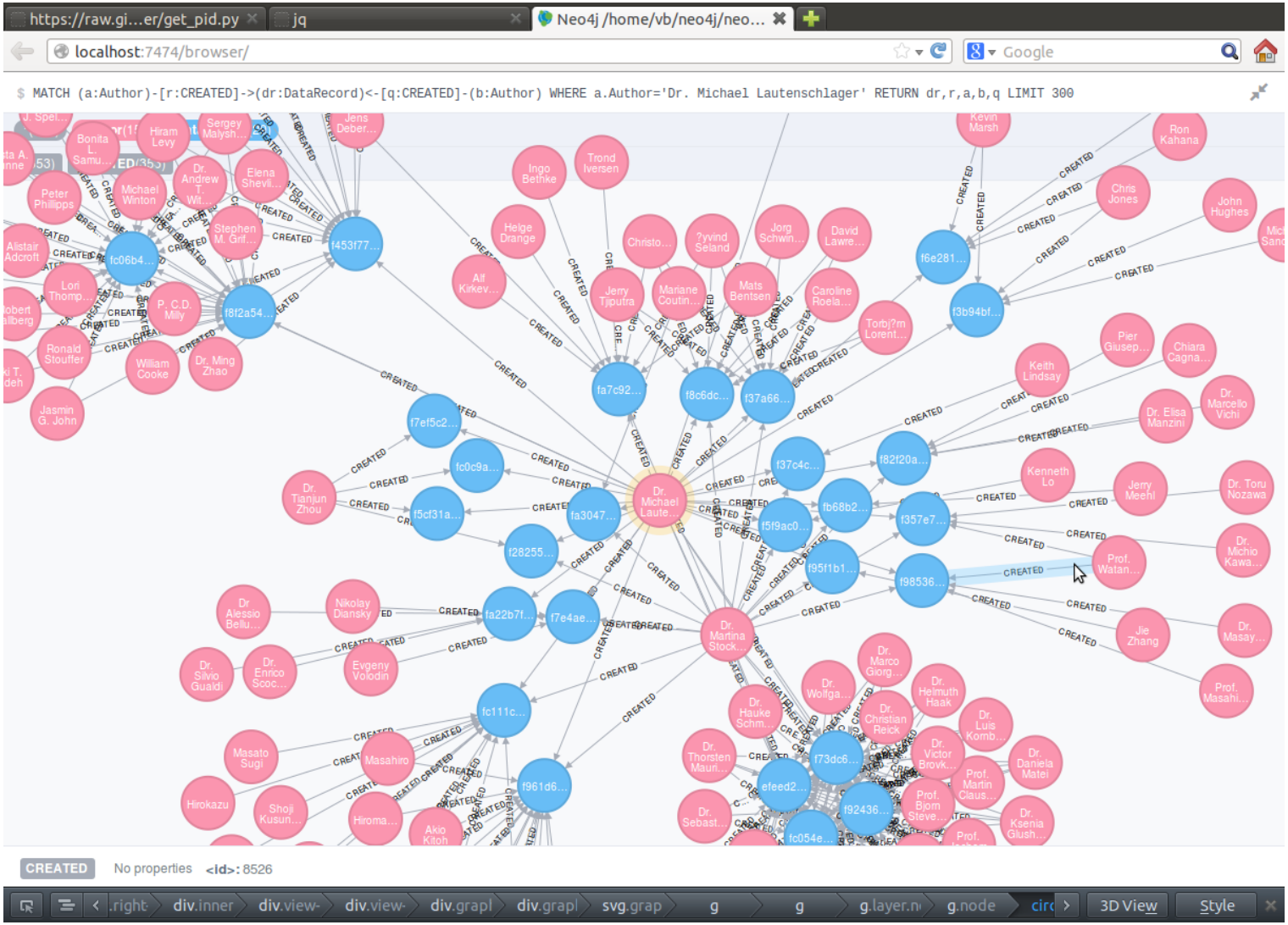
Plain/raw B2FIND record uploaded in neo4j.
Take a note of multiple Authors squeezed in one field, and the same applies to Publishers



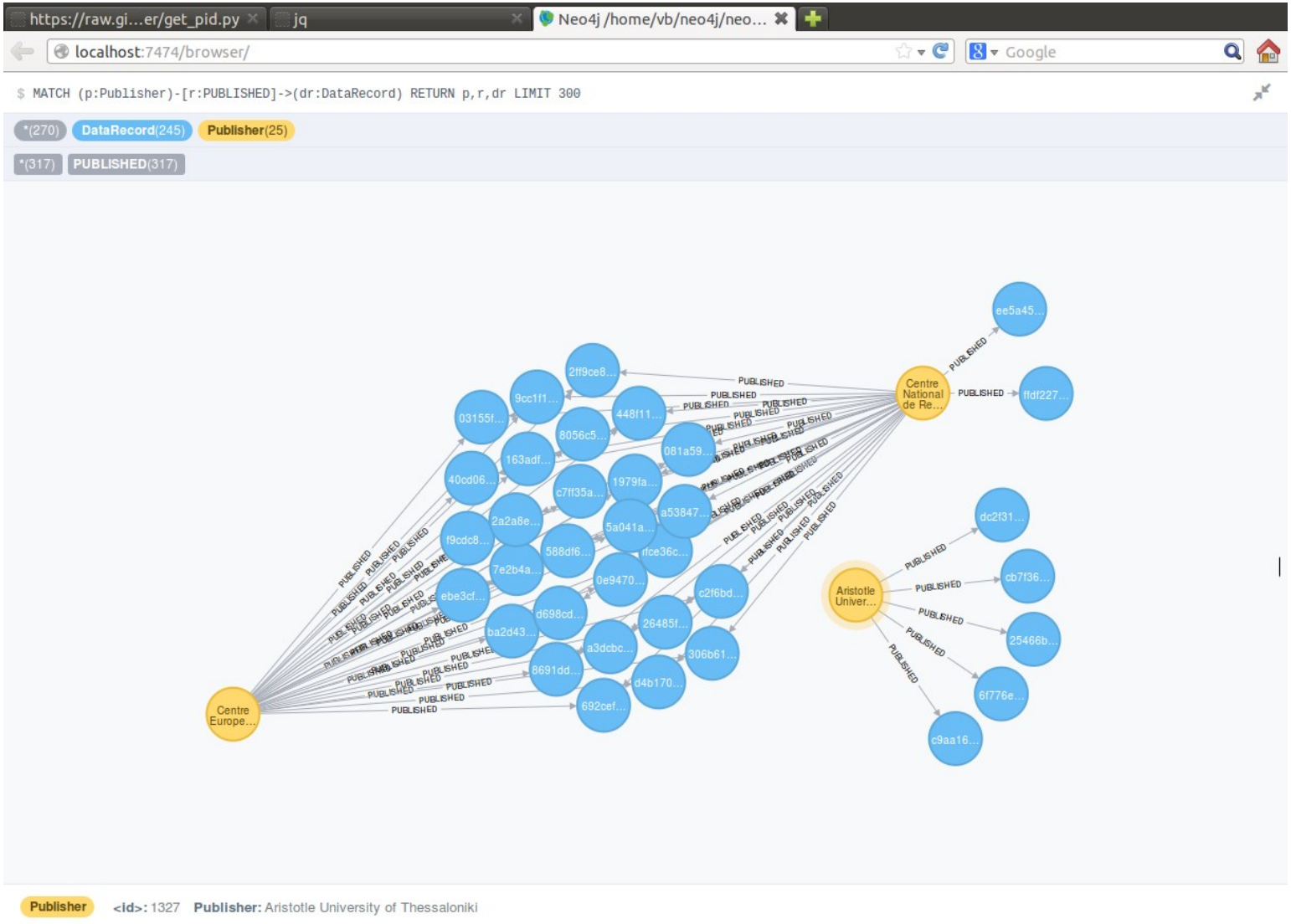
Many benefits should come from conversion of a raw B2FIND record into the graph model with new nodes created for Authors, Publishers, Languages and Tags



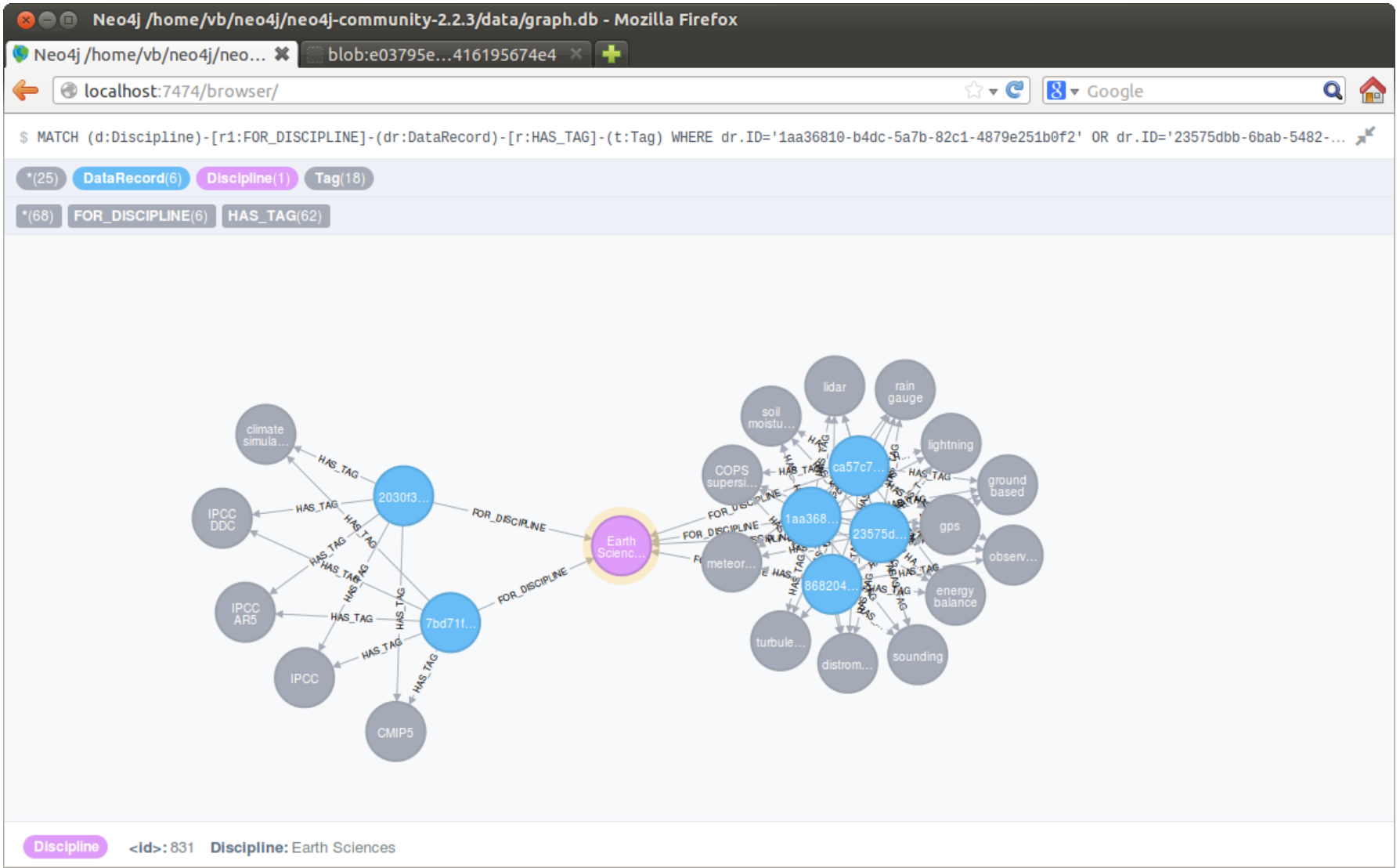
Authors (data creators) become new nodes. This allows insights in data co-authorship and contributes to finding related data



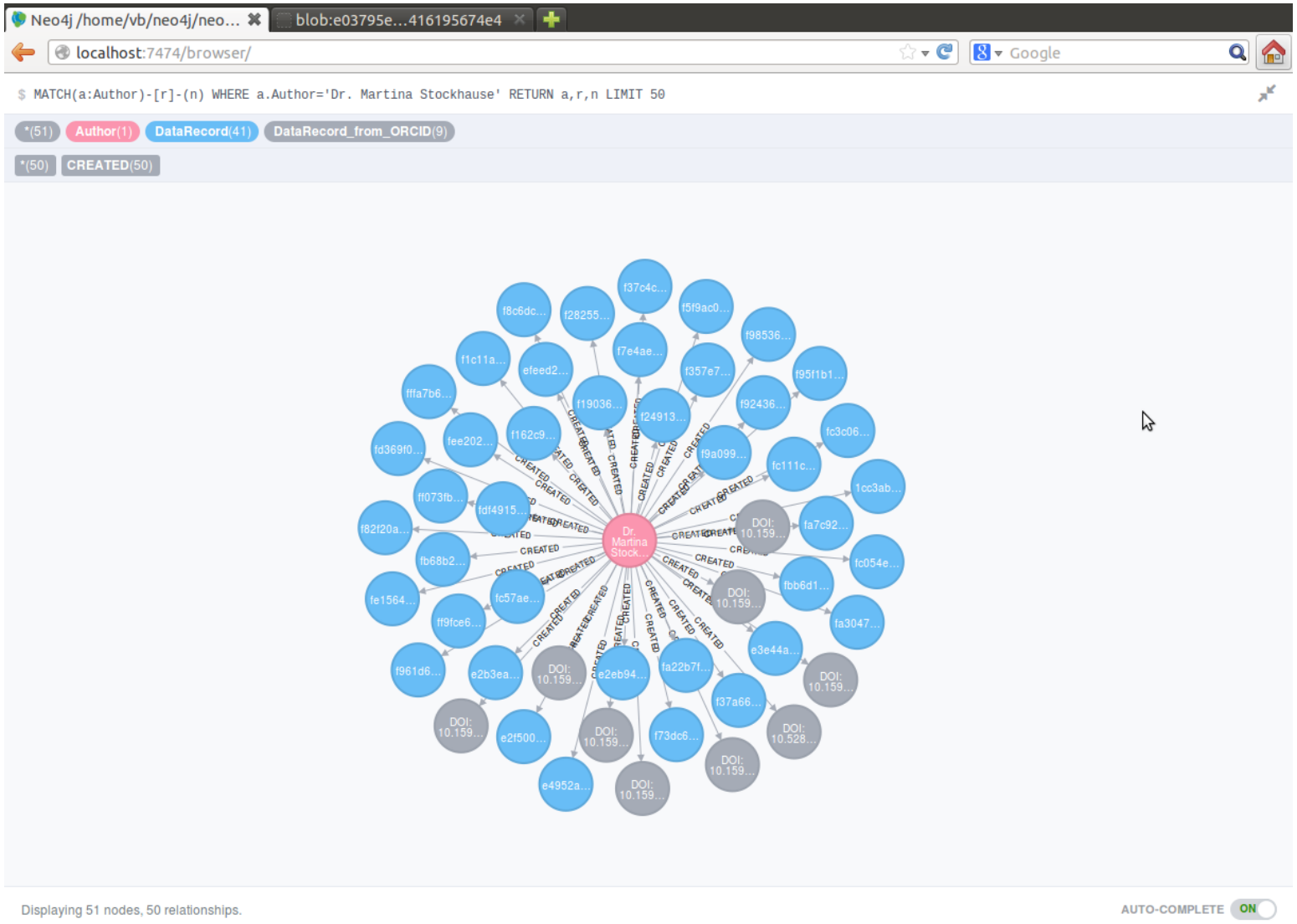
Data publishers become nodes.
This allows insights in collaboration between organizations



Tags become nodes. This allows looking for topical clusters and can support new flavours of data search, like “find heavily (or less heavily) related data”



Experiment with ORCID records (for one selected Author): it is possible to mix up differently structured data records in one graph



Why (a small scale) experiment with ORCID records can be of interest

- ORCID records have differently structured metadata yet can be successfully mixed up with B2FIND records in one graph, then queried as necessary
- It turned out that none of data records obtained from ORCID were harvested by B2FIND, so ORCID can complement current B2FIND data sources
- Beyond ORCID, there are other perspective data sources to explore, e.g. Figshare
- In short, B2FIND now covers institutional repositories (and does this fairly well) yet ORCID and Figshare are examples of repositories popular with individual researchers; they are the "long tail of data records" that B2FIND may want to cover, too, to become a truly universal research data catalogue

Thank you!

data
LHC
research
e-Science
particle
petabytes
used
Tier-1
first
November
energy
physics
services
collisions
experiments
experimental
STFC
R89
new
also
year
users
construction
researchers
rates
UK
number
years
Rutherford
Appleton
Laboratory
computing
processing
universe
machine
CERN
March
impact
handling
protons
including
high
Two
TeV
distributed
experiment
activities
plasma
smoothly
sites
support
economic
time
requirements
simulating
initiatives
room
beams
move
start
network
Research
events
CMS
LHCb
SCARF
centre
world
Atlas
expected
storing
programme
analysis
last
project
recorded
science
started
received
main
store
four
tier-1