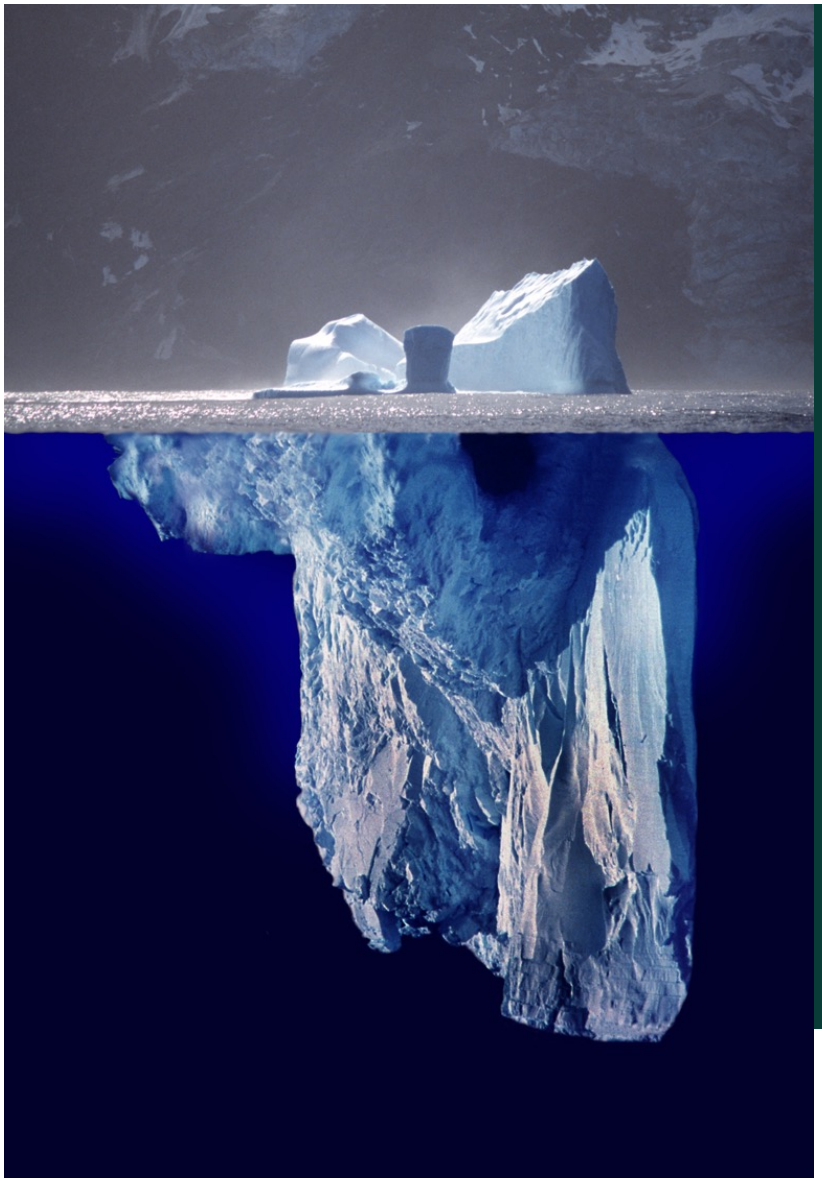
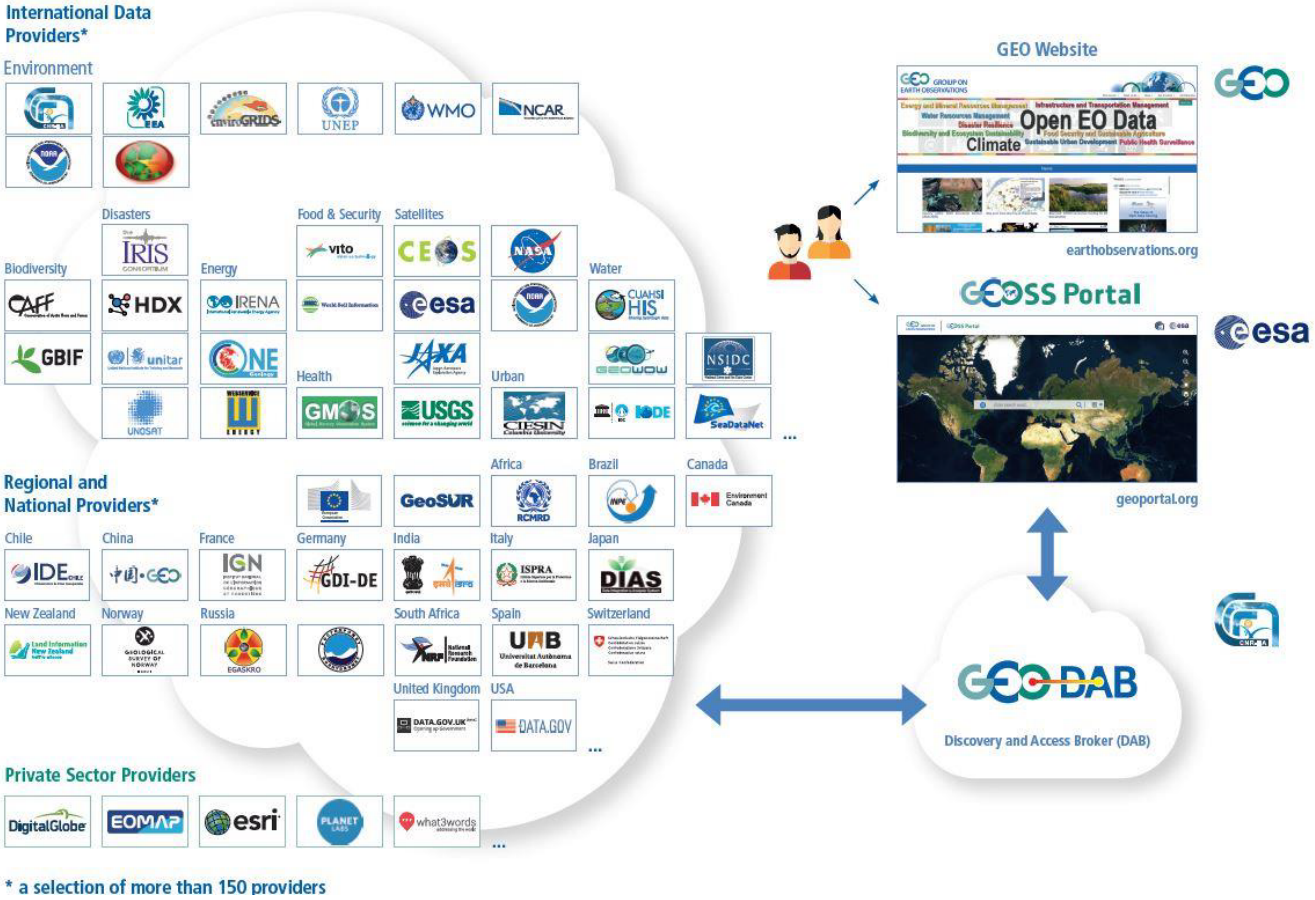


Exploring the depths of the GCI

Jiri Hradec , Max Craglia, (JRC)
Mattia Santoro, Stefano Nativi (CNR)

European Commission

We know about the GCI….



Some Key Numbers:
About 45 million datasets;
More than 200 million granules

European Commission

but do we really?

- As far as we know nobody has ever looked inside the GCI to see what's there.
- Exploratory project by JRC and CNR, building on a similar analysis done on the INSPIRE geoportal.
- We have just started scratching the surface, so these are very preliminary findings, warts and all…

The What

- Analysed 1.8 million metadata records harvested by the GEODAB.
- These include all the large data collections except GIBIF and HIS (in-situ hydrological data)
- As the MD of the collection represents well the MD of the siblings: e.g. CEOS has about 4300 MD records of collections and about 286M MD records of individual scenes which are not harvested, the MD used represents more than 85% of all data in GEOSS.

The How

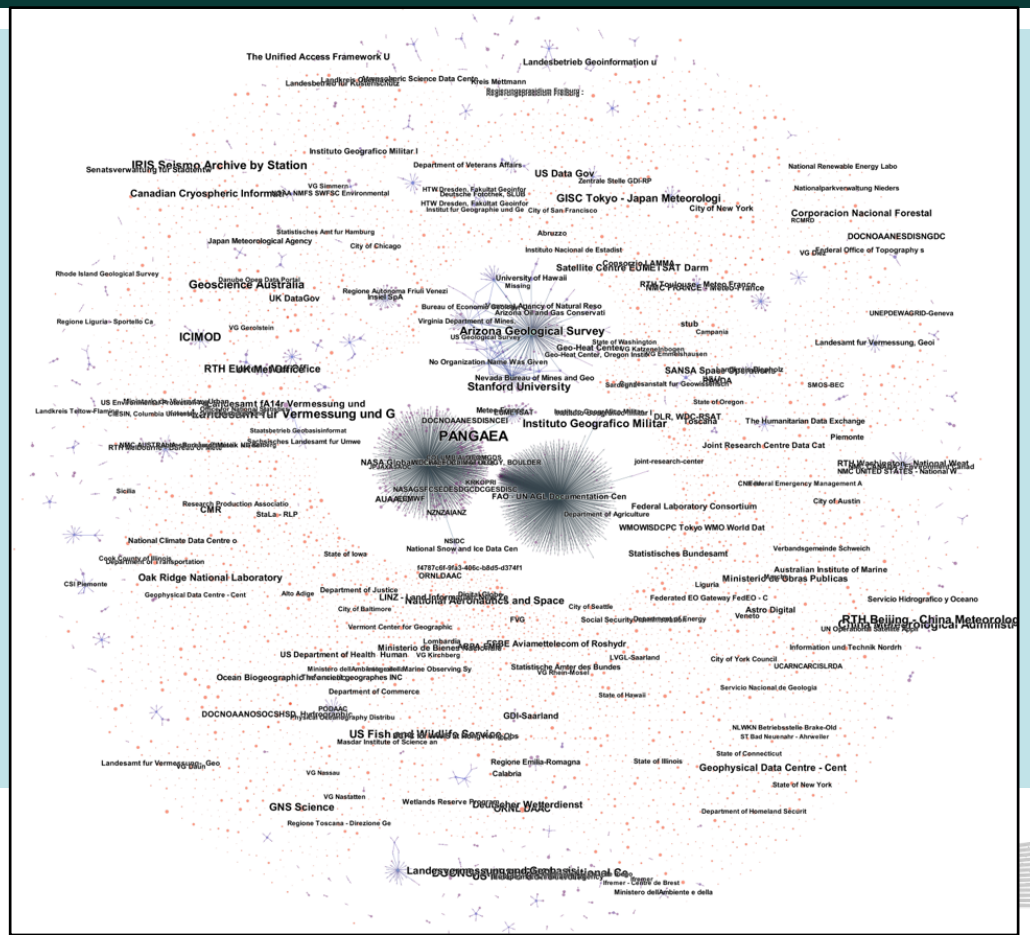
- Analysed each record for language IDs
- Machine translated all records not in English
- Parsed the content of the MD records, analysed counts, frequencies, and semantic relations among the 195,000 keywords, and text in title and abstracts using neural networks.

Who are the providers?

- A very good question….
  - Sometimes difficult to distinguish between roles – data provider, data distributor, data center, contributor, author etc. Meanings of these key words differ greatly among data hubs.
  - There seem to be more than 3500 data providers (originators) but some are organisations, others are data hubs with other organisations inside.
  - More than 10,000 contact orgs but many thousands are the URL of individual images!
  - 5-10% of institutions do not exist but we do not know whether it is because the institution has changed name or because it is archived data from the past.

Step 1: connect providers

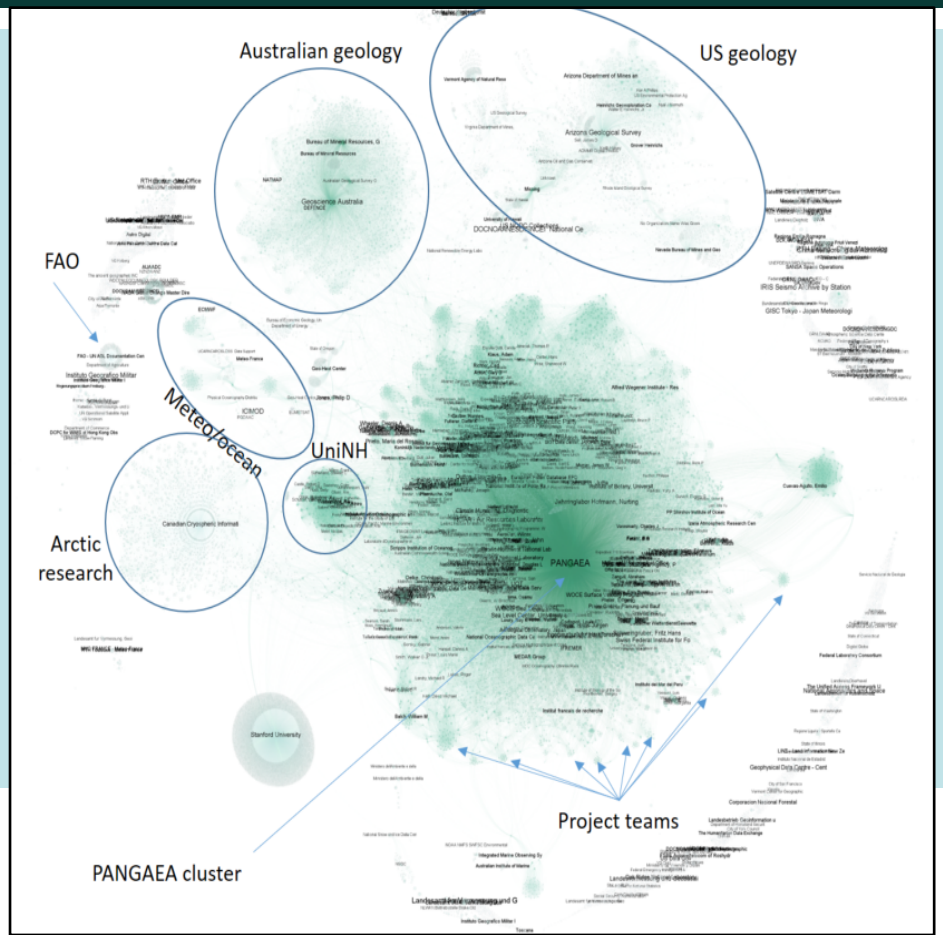
- Connected Contact Org. and Distributor



European Commission

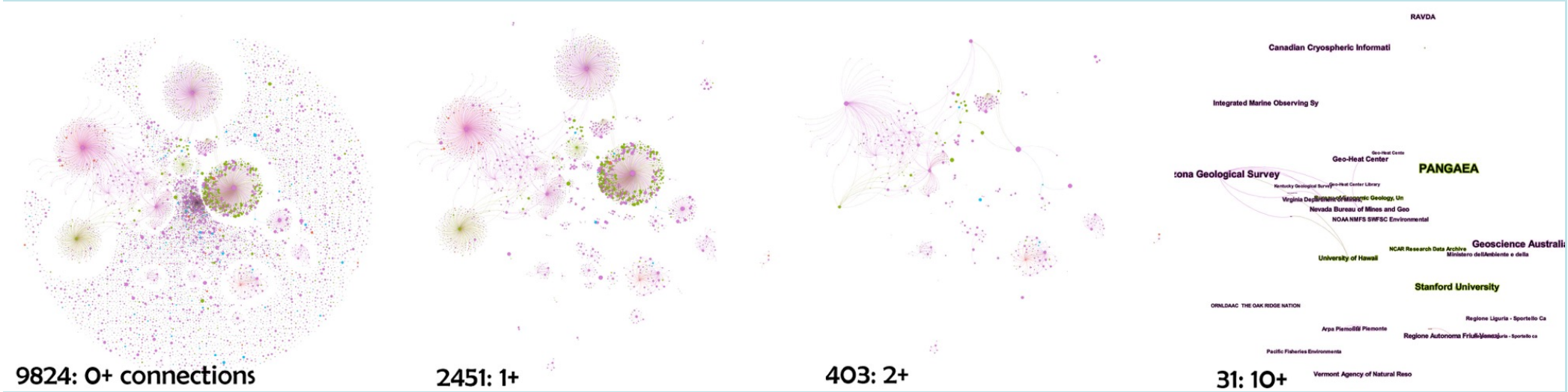
Step 2: citations among organisations

- 66,000 nodes
- 350,000 edges



European Commission

Step 3: geocoded and filtered by strength of connectedness



European Commission

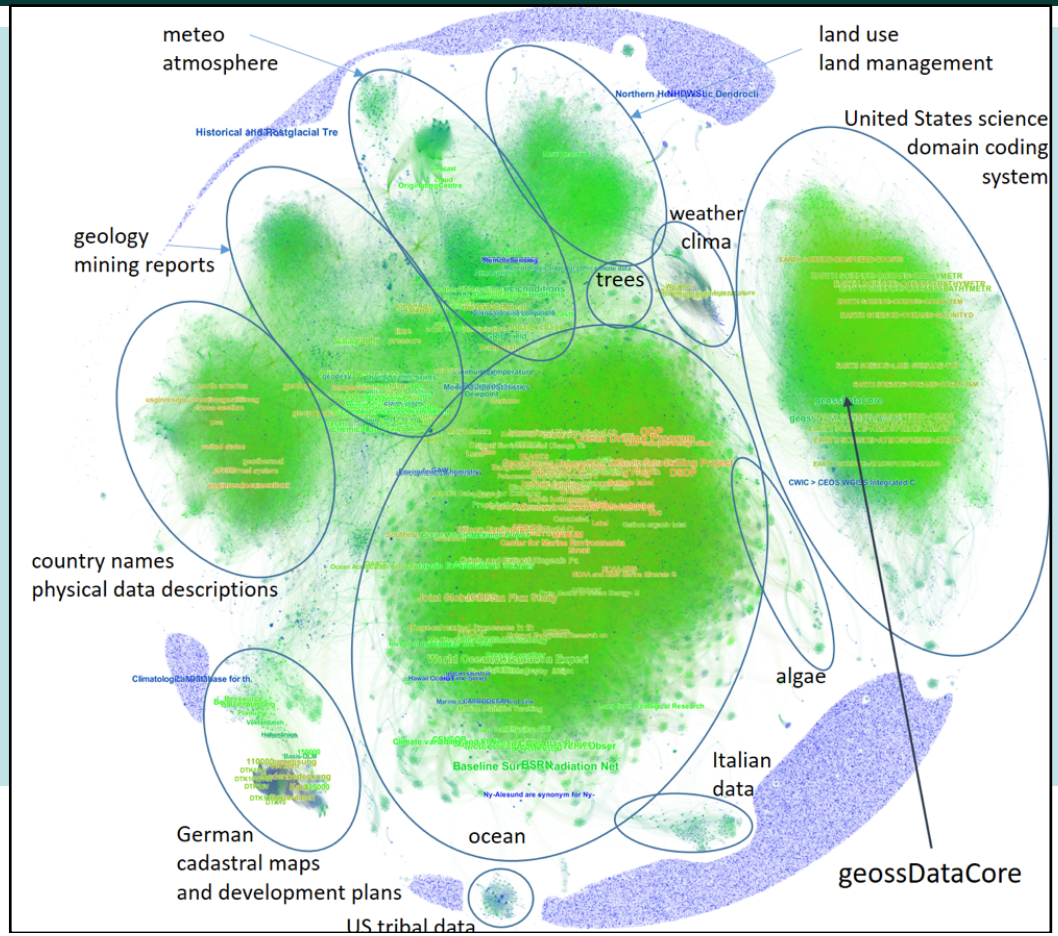
Thematic Coverage

- At least 1 keyword in each MD record up to 666> total 198,000 keywords
- Huge variety as we lack a shared Thesaurus to help describe the most common concepts.
- Abstracts are the richest source of knowledge

10 most frequent keywords

Baseline Surface Radiation Network BSRN)	132,785
World Ocean Circulation Experiment (WOCE)	59,766
Ocean Drilling Program (ODP)	45,707
Historical and Postglacial Tree Ring Archive of Hohenheim (HISTRA)	41,858
Joint Global Ocean Flux Study (JGOFS)	28,819
Deep Sea Drilling Project (DSDP)	21,890
Meteorological Long-Term Observations @ AWI	20,791
Northern Hemispheric Dendroclimatological Network	16,259
geossNoMonetaryCharge	14,976
geossDataCore	14,915

Keywords clustering

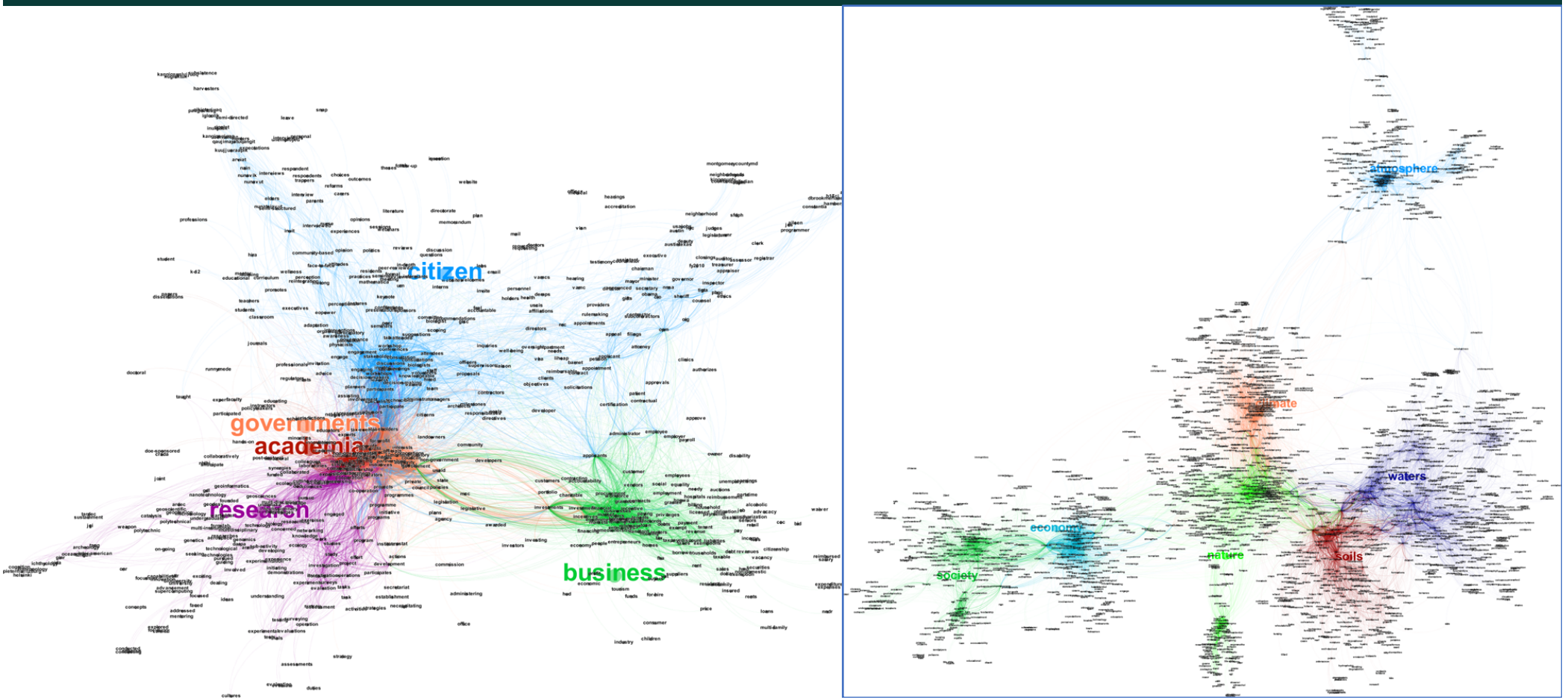


European Commission

Neural network analysis Abstracts and Titles

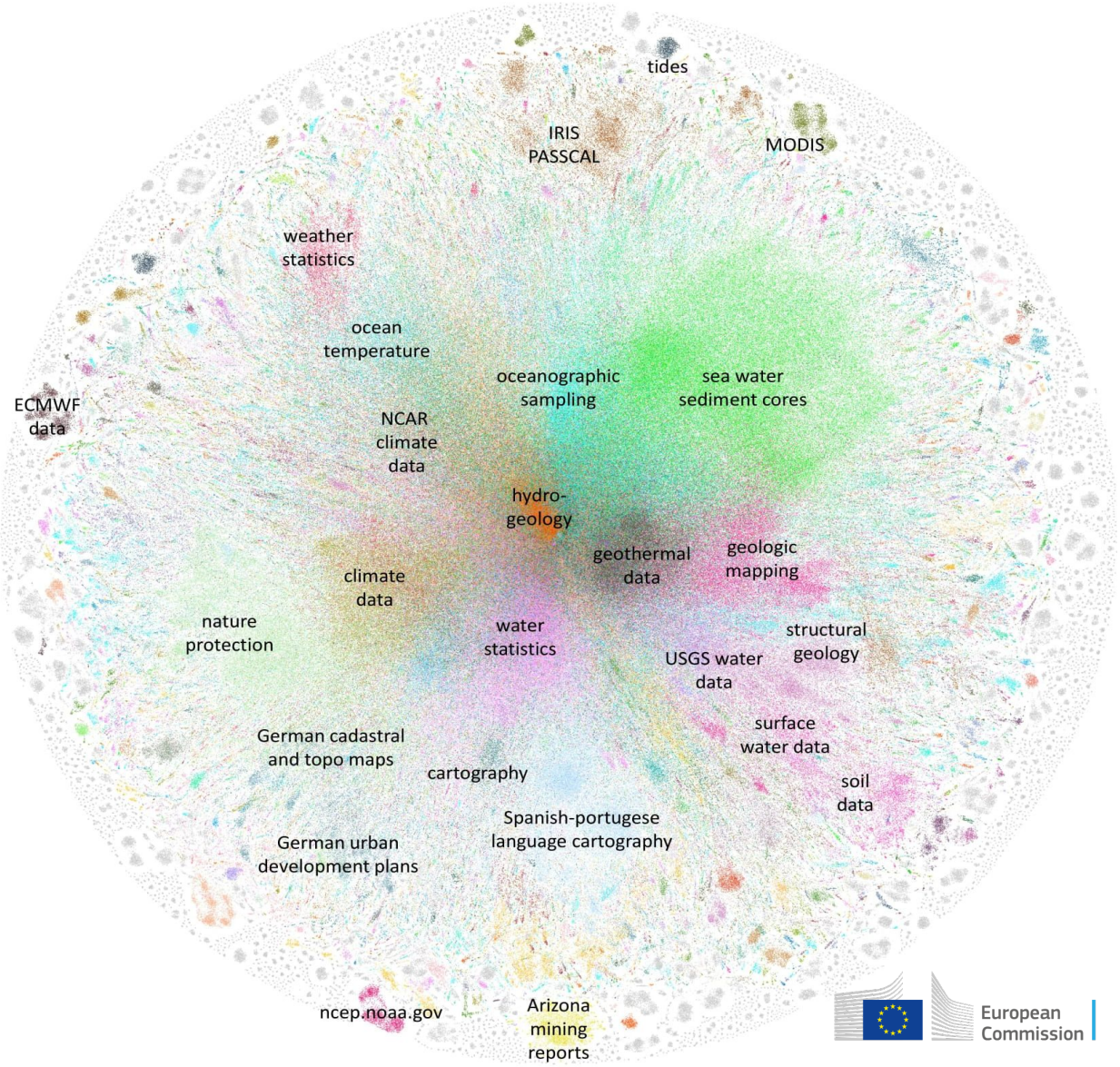
- 1.8 M records
- 653 M lemmatized words
- CBOW model to extract higher level concepts and relationships

Semantic relationships in the abstracts



GEOSS data universe

Neural network analysis by the European Commission Joint Research Centre and the Italian National Research Council of the 635 million words in the titles and abstracts of 1.8 million metadata records in GEOSS (representing 89% of all GEOSS data)



Conclusions

- To our knowledge first ever analysis of global system of system of this size and complexity
- Given its vary nature as System of Systems heterogeneity is to be expected
- Difficult to parse complex xml into flat formats (e.g. tables for statistical analysis)
- The first level analysis shows some problems such as lack of persistent IDs, difficulty in understanding which is data is well curated and maintained, and which is is not, and which data is authoritative and can be relied upon, and which is not.

Conclusions

- There are also good news e.g. most data is full and open access, and we can identify paths for improvement such as shared thesaury or list of common concepts, cross domain mappings, improved data management practices, including persistent IDs.
- We have also identified the most connected institutions to work with as a priority
- Stay tuned as we keep digging!

European Commission

Any questions?

Massimo.craglia@ec.europa.eu

European Commission