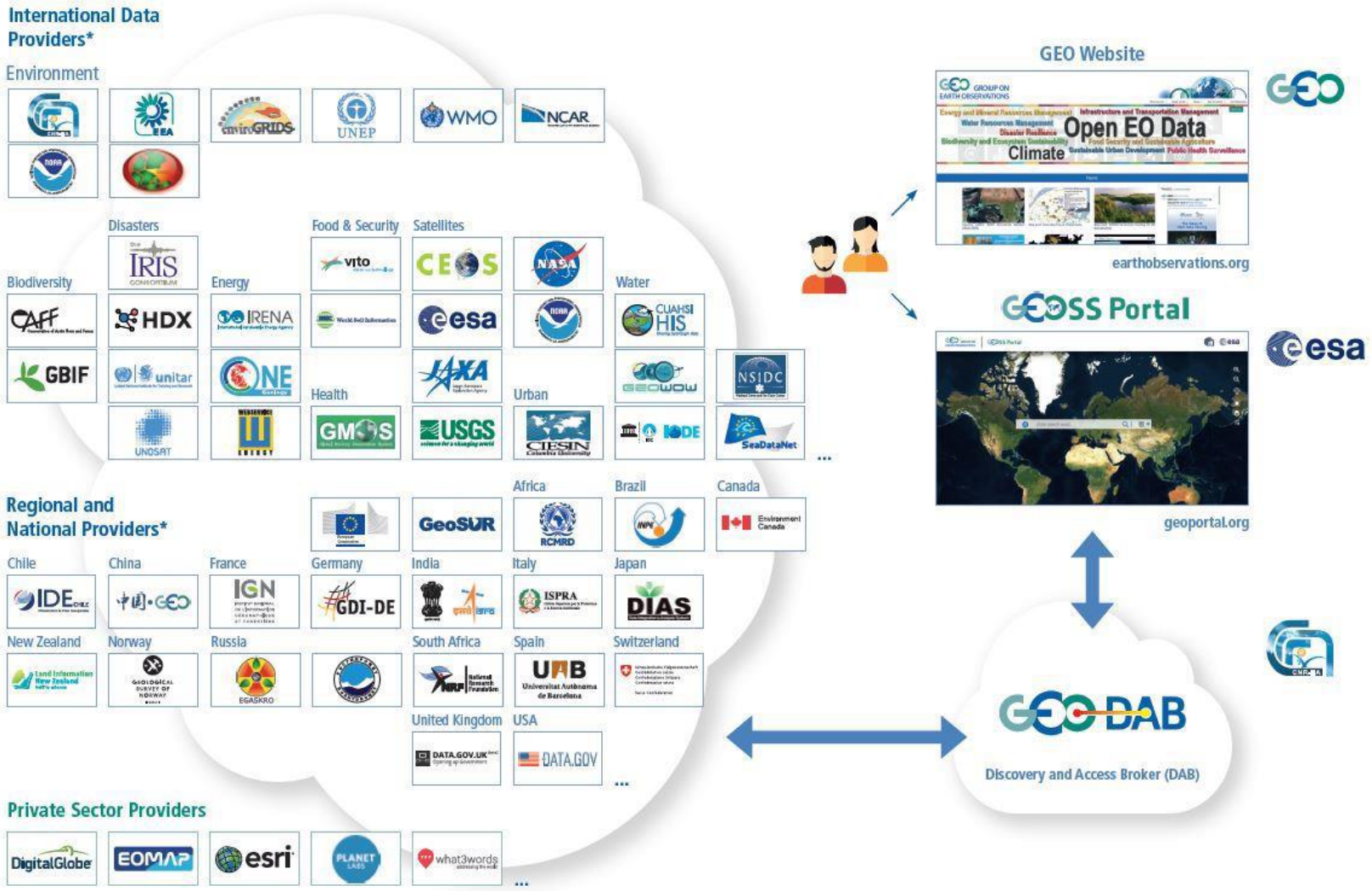
Data Sharing in GEOSS

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 European Commission Joint research Centre and Italian National Research Council.
- Analysed 1.8 million metadata records harvested by the GEODAB.
- These include, more or less, all the large data collections except GIBIF and HIS (in-situ hydrological data)
- All together the analysis covers 89% of all data in GEOSS

The How

- Machine translated all records not in English (35 languages found)
- Parsed the content of the MD records, analysed counts, frequencies, and semantic relations among the 195,000 keywords
- The abstracts proved the most valuable source of content. We created a corpus of 635 million words from all the titles and abstracts, and then analysed with artificial neural networks.

Use conditions

- Out of 1.8 M records, 1.4 M are geossdatacore or no constraints
- Open with limited constraints about 100k
- Quite a lot of variety of terms and conditions
- Some 100k records point to explicit constraints (often mapping agencies)
- Some 200k records have no clear indication.
- 520 ways to say I am open!

use_constraints: Descending	Count
geossdatacore	355,237
License not specified	45,817
Distribution liability: NOAA and NCEI make no warranty, expressed or implied, regarding these data, nor does the fact of distribution constitute such a warranty. NOAA and NCEI cannot assume liability for any d	31,064
Geobasisdaten (analog und digital) sind gesetzlich geschÃ¼tzt. Unbefugte VervielfÃ¤ltigung oder Verbreitung verstÃ¶Ã§t gegen das Urheberrechtsgesetz (UrhG) und das Schleswig-Holsteinische Vermessungs-	26,995
Die Geobasisinformationen sind nach Urheberrechtsgesetz (Datenbankhersteller) und nach dem Gesetz Ã¼ber die Landesvermessung und das Liegenschaftskataster (Vermessungs- und Katastergesetz - VermK	25,885
U.S. Government Work	16,585
GTS Priority 3	14,842
S/N	13,183
WMO Additional	11,775
[]	11,276
(C) Institute of Geological and Nuclear Sciences Limited.	9,690
Free to use with copyright to ICIMOD. Citation: ICIMOD (2014) MODIS processed snow data. Kathmandu: ICIMOD, (http://rds.icimod.org).	9,547
UK Open Government Licence (OGL)	8,453
Other License Specified	7,521
http://www.geobasis-bb.de/pdf-Dateien/AGNB.pdf	6,950
This material is licensed under a Creative Commons Attribution 3.0 New Zealand (CC BY 3.0) Licence. For more details visit http://creativecommons.org/licenses/by/3.0/nz. Where the data are used in a figure	6,164
Wenn Sie Informationen bzgl. der Nutzungsbedingungen und anfallender Kosten brauchen, wenden Sie sich bitte an die angegebene Kontaktstelle.	5,819
WMOEssential	5,544
WMOAdditional	4,882
Geobasisdaten (analog und digital) sind gesetzlich geschÃ¼tzt. Unbefugte VervielfÃ¤ltigung oder Verbreitung verstÃ¶Ã§t gegen das Urheberrechtsgesetz (UrhG) und das Schleswig-Holsteinische	4,151
This material is licensed under a Creative Commons Attribution 3.0 New Zealand (CC BY 3.0) Licence. For more details visit http://creativecommons.org/licenses/by/3.0/nz. Where the data are used in a figure	3,525
Free to use with copyright to ICIMOD. Citation: ICIMOD (2014) MODIS processed aerosol data. Kathmandu: ICIMOD, (http://rds.icimod.org).	3,481
WMO Essential	3,238
Creative Commons CCZero	2,781
keine	2,527
hdx-other	2,066
no conditions apply	2,040
Free to use with copyright to ICIMOD. Citation: ICIMOD (2014) MODIS processed cloud data. Kathmandu: ICIMOD, (http://rds.icimod.org).	1,972
WMOOther	1,915
http://www.geobasis-bb.de/GeoPortal1/pdf/AGNB.pdf	1,761
Nessuna	1,596
Data, products and services from IMOS are provided "as is" without any warranty as to fitness for a particular purpose.	1,556
None	1,446
Sin restricciÃ³n	1,406
(c) Statistisches Bundesamt, Wiesbaden 2016, Genesis-Online; Datenlizenz by-2-0	1,404
small scale application only, map scale 1:250,000 to 1:4,000,000	1,403

Processing fees

- Out of 1.8 M records only about 50k explicitly require fee to access
- When fee requested a lot of confusion on costs, units of measurement, and terms for calculation

processing_fees: Descending	Count
Digital data may be downloaded from NCEI at no charge in most cases. For custom orders of digital data or to obtain a copy of analog materials, please contact NCEI Information Services for information al	31,067
Se debe contactar con la oficina de Marketing al 593-2-3975169 o por email: igm@mail.igm.gob.ec, haciendo referencia a asunto: Senores Marketing.	13,916
Electronic download of the data is free. There is a fee for copies of data on physical media.	9,689
kostenfrei	5,919
8.41	4,916
0	2,017
20	1,479
5.00 Euro	1,473
0	1,436
4.50 Euro	1,422
6.00 Euro pro km2, 15.00 Euro Mindestbestellwert	1,083
4.00 Euro	1,051
7.50 Euro pro kmÂ²; Mindestbestellwert: 15.00 Euro; HÃ¶chstentgelt: 55 000.00 Euro	1,040
4.00 Euro pro km2, 15.00 Euro Mindestbestellwert	1,014
nach GebÃ¼hrensatzung	570
90	550
keine	521
Informacion Liberada, excepto la Cartografia Reservada	512
P	371
7.00 Euro	328
1.00 Euro pro km2, 15.00 Euro Mindestbestellwert	292
0.75 Euro pro km2, 15.00 Euro Mindestbestellwert	285
N	245
geldleistungsfrei	211
29.5	181
Datenabgabe gemÃ¤ÃŸ Umweltinformationsgesetz (UIG) und NiedersÃ¤chsischer Allgemeiner GebÃ¼hrenordnung (AllGO), OpenData nach Datenlizenz Deutschland Namensnennung	163
Es werden GebÃ¼hren entsprechend der Kostenordnung fÃ¼r das amtliche Vermessungswesen erhoben (letzte Fassung: Nds. GVBl, Nr. 2/2006 v. 19.01.2006).	134
14,02 EUR (zzgl. 7% MWSt. und Versandkosten)	121
bei DatentrÃ¤gerabgabe gemÃ¤ÃŸ GebÃ¼hrenordnung	118
kostenlos bzw. nach Bereitstellungsaufwand	118
fÃ¼r Papierabgabe nach GebÃ¼hrensatzung	103
fÃ¼r Berechtigte Nutzer kostenfrei, bzw. nach Bereitstellungsaufwand	101
9	88
0.30 Euro pro km2, 15.00 Euro Mindestbestellwert	87
23,36 EUR (zzgl. 7% MWSt. und Versandkosten)	87
0.25 Euro pro km2, 15.00 Euro Mindestbestellwert	86
Kostenordnung fÃ¼r das amtliche Vermessungswesen (KOVerm): http://www.nds-voris.de/jportal/?quelle=jlink&query=VermKostO+ND&psml=bsvorisprod.psml&max=true	86
none	84

Data Services: Ouch! From dead services to unknown formats

digital_forms.name: Descending	Count
(none specified)	1284402
text/tab-separated-values	528119
MiniSEED	501024
text/html	43016
application/x-pdf	37886
application/xml	26961
image/png	25180
HINWEIS: Die Abgabe der digitalen Daten ist kostenpflichtig. Lieferung: digital, CD, DVD, TIFF unkomprimiert. Analoge Daten: plot-on-Demand.	21740
image/jpeg	21377
GRIB1	21008
text/csv	18792
image/tiff	18769
application/zip	18416
ASCII, HINWEIS: Die Abgabe der digitalen Daten ist kostenpflichtig.	18281
GRIB	18150
image/gif	14926
image/svg+xml	14242
pdf	13062
TIF	12978
analogico	12933
image/png; mode=8bit	12256
application/json	11120

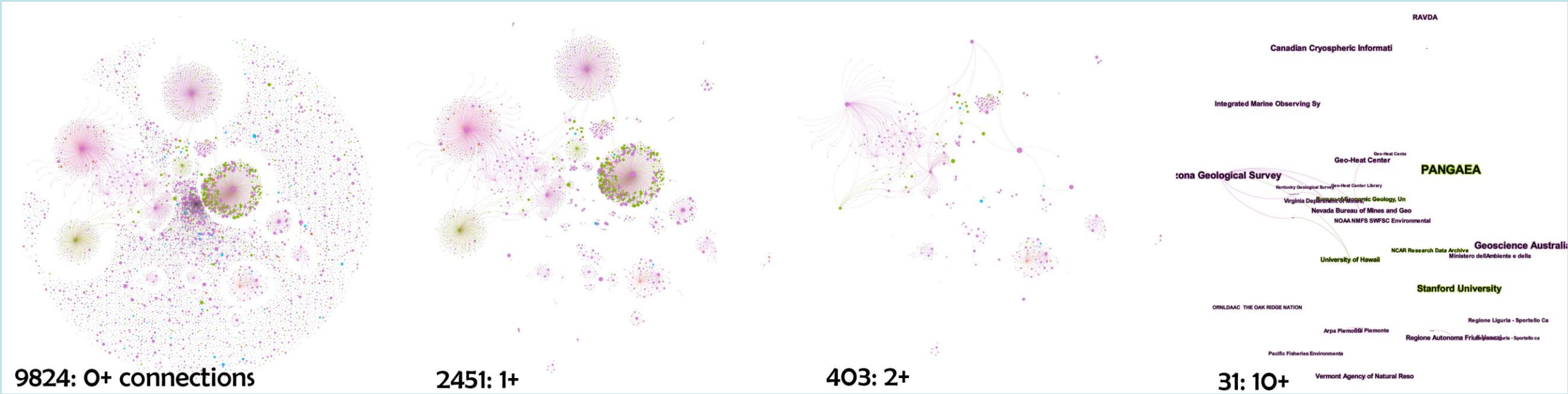
POSITIVE
They exist (sometimes)
They actually allow user to download some data

NEGATIVE
25% of services are at IP addresses / domain names that do not exist
30% of the rest is not reachable at the URL specified
Data format often none or meaningless for computer use.
Except for MiniSEED and CSV files, formats point to mostly text files like PDF instead of actual data

European Commission

How to improve: work with key data providers

Connectedness among data providers in GEOSS



European Commission

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
- 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, there is metadata, 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, and consolidation at the level of major facilities/collections.
- Stay tuned as we keep digging!

Any questions?

Massimo.craglia@ec.europa.eu

European Commission