Data management and sharing

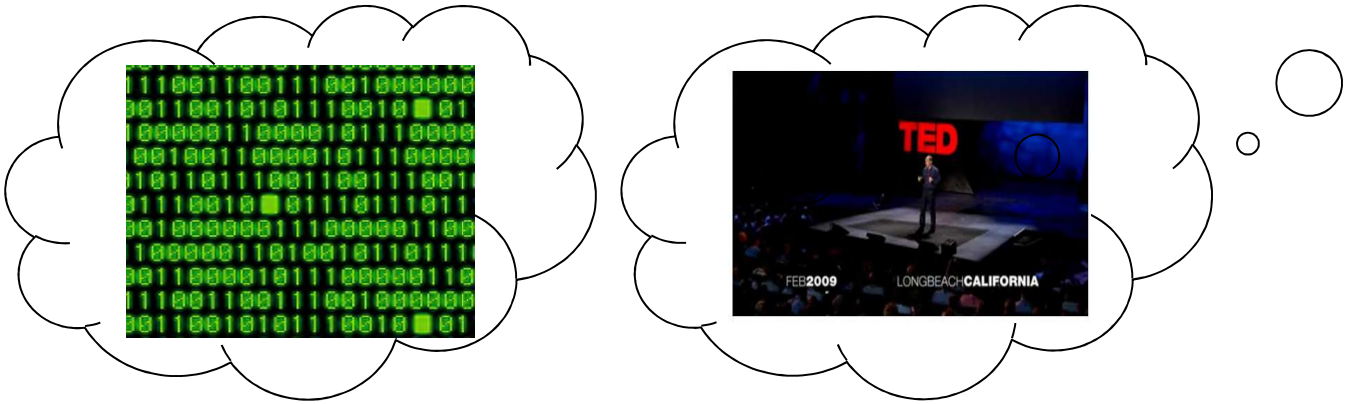
Dr. Gerben J. de Boer
Gerben.deboer@vanoord.com

- Van Oord: data scientist
- Tu Delft: visiting scientist
- MODEG: member 2010-2015

22 jun 2017
JERICO-next
summerschool

Standards: should be 100% cloud

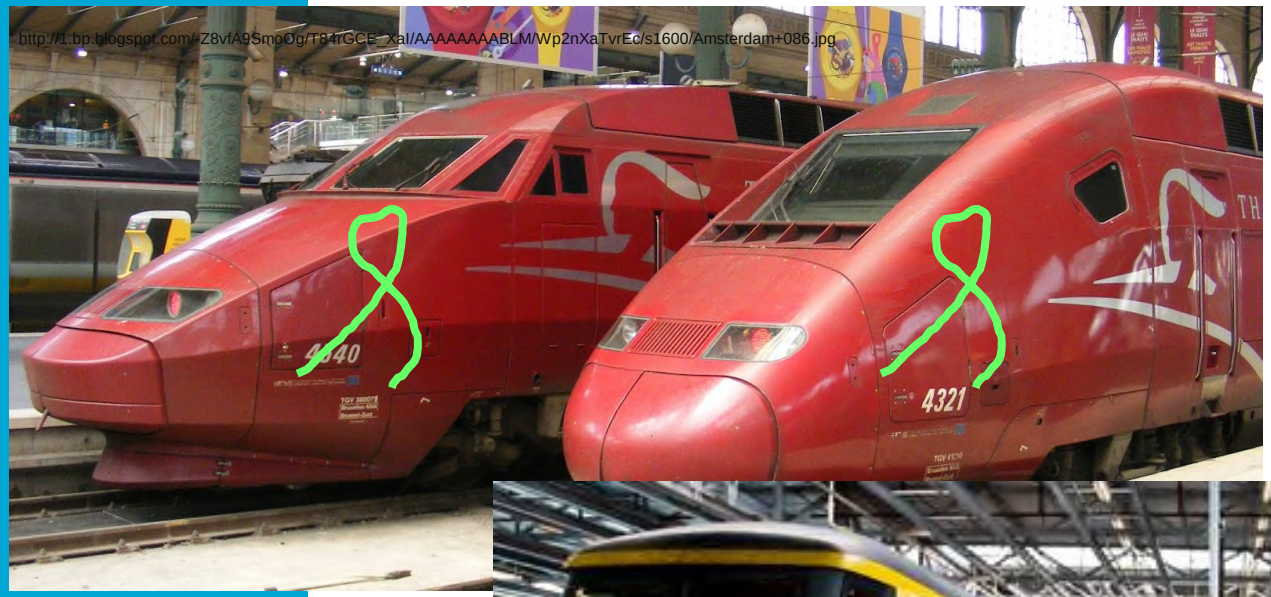
Different layers of standard inspired by www inventor Tim Berners-Lee: **5 ★ Open Data**



TU Delft

Van Oord
Marine ingenuity

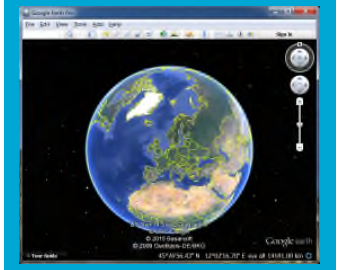
Standards: ***existing***

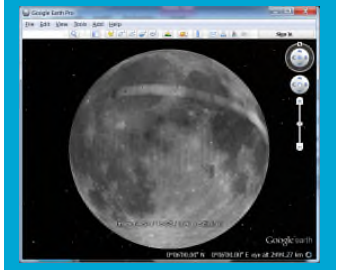


Use proven and existing technology, avoid being Guinee pig for too new standards. netCDF: made first in 1989. OGC standard in 2013.

The Dutch ordered hispeed trains using the latest EU ERTMS standard. It failed and the trains were returned. The Dutch did not use older standards as back-up, even thought these older standards already work. The Thalys uses a range of standards.

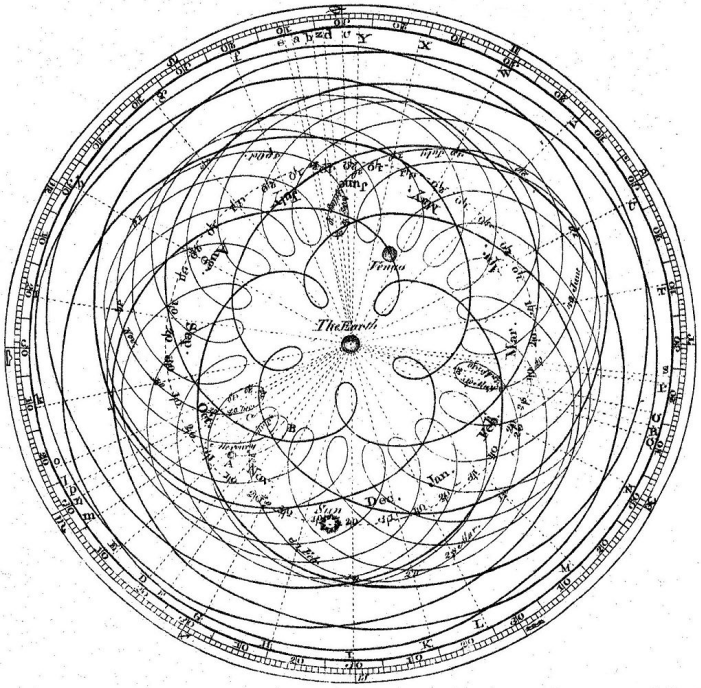
Google Earth can even handle multiple **planets**. Who records the planet as meta-data?

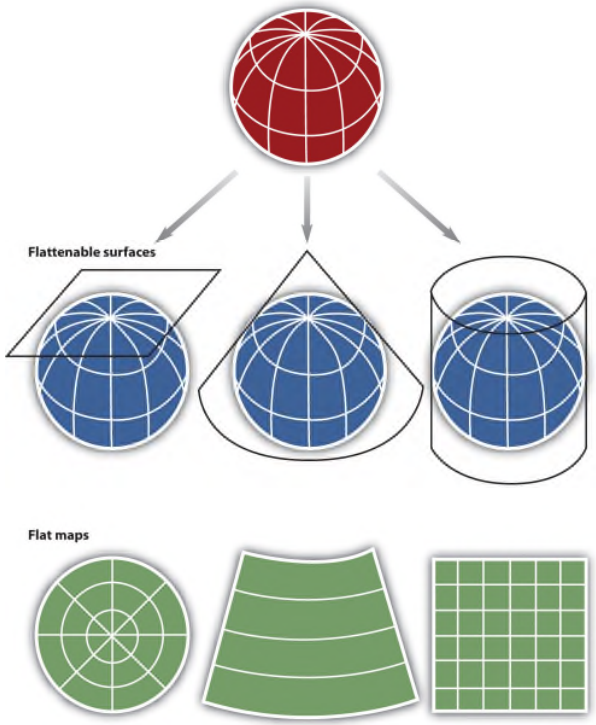




Standards: *1 earth, 1 coordinate system*

Google Earth overtook all GIS sciences and standards with Google Maps and open 3D kml: use only GPS WGS84. And UTC for times.



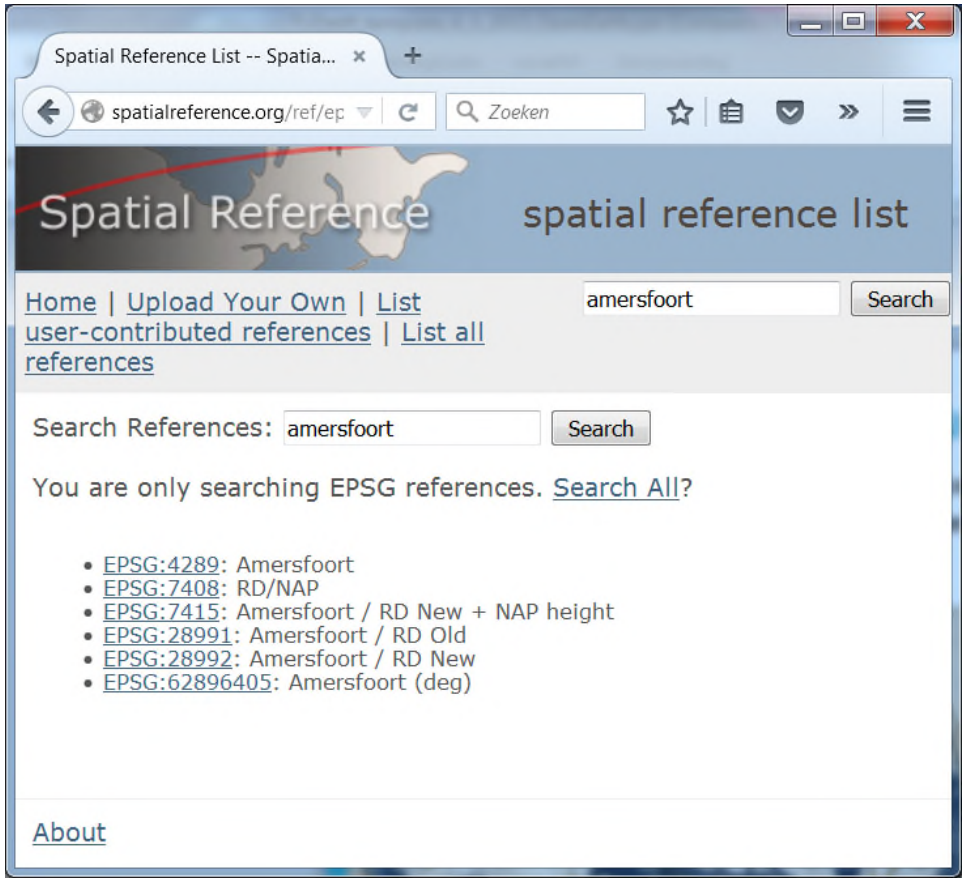


https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s06-map-anatomy.html

Enormous complexity of an earth-centered system became simple with a sun-centered one. Occams' razor at best.

Coordinates become easy when you ignore country-based earth systems that try to make earth flat. Use a global system for a *sphere*.

Standards: *flat earth standards*



If you are forced to use a flat-earth coordinates, pick one from the global EPSG list.

The EPSG code database made first by oil & gas industry.

Only later is was adopted as a standard, because it works.

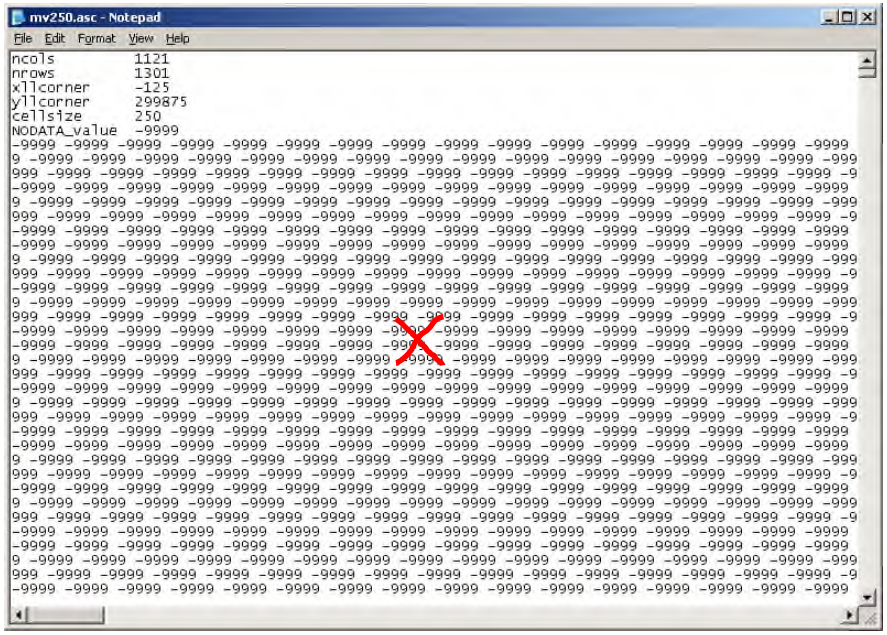
Still in 2015 the Dutch government has 5 ways to map coordinates. Which one to choose ? NONE. Use GPS + WGS84. The Earth is round !

Van Oord
Marine ingenuity

TU Delft

3 ✗
3N ✗
F=3 ✗
F = 3N ✓

Standards: ***metadata***



ESRI ArcGIS ASCII "standard" would fail at high school physics exam.

TU Delft

Van Oord
Marine ingenuity

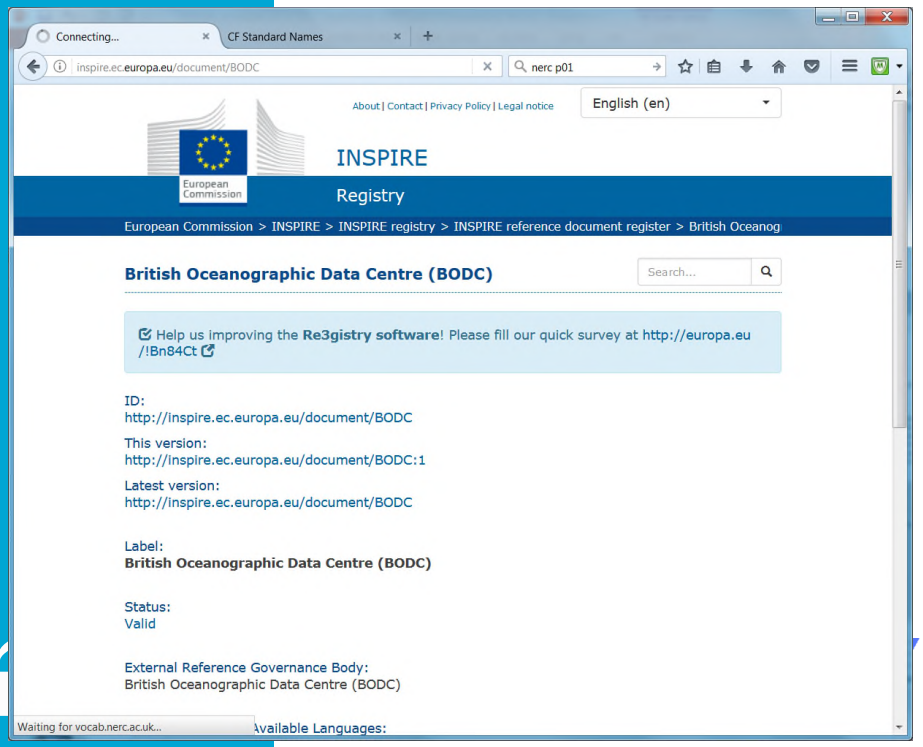
3 ✗
3N ✗
F=3 ✗
F = 3N ✓

Standards: quantities (F), units (N)

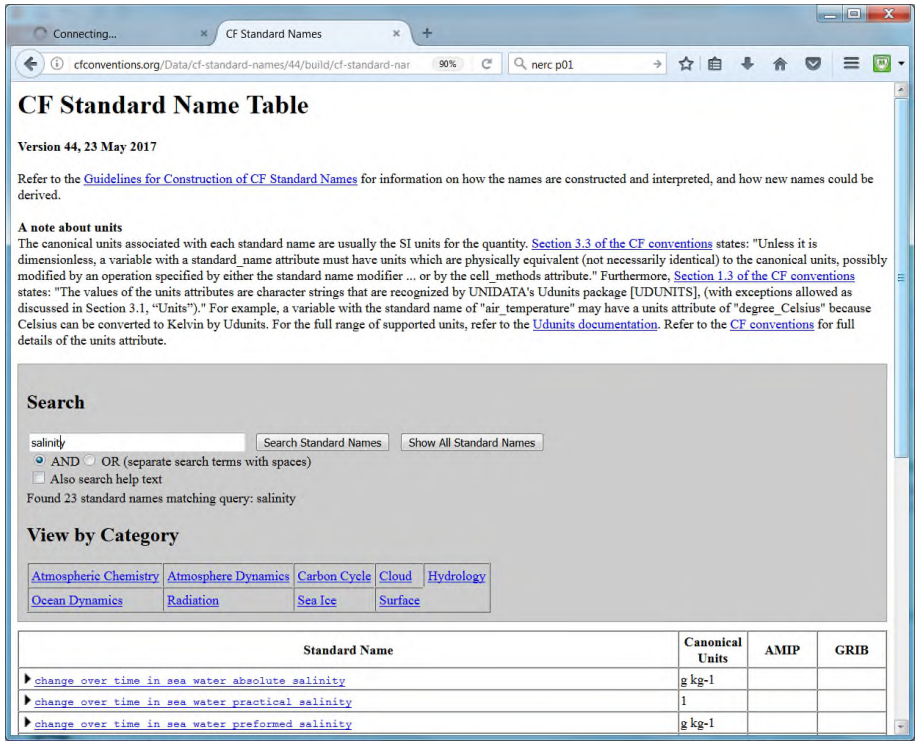
Use quantities and units from internationally agreed-upon lists.

Chemicals: CAS
Species: WoRMS
Quantities: P01 or CF

NERC P01



CF standard names + units

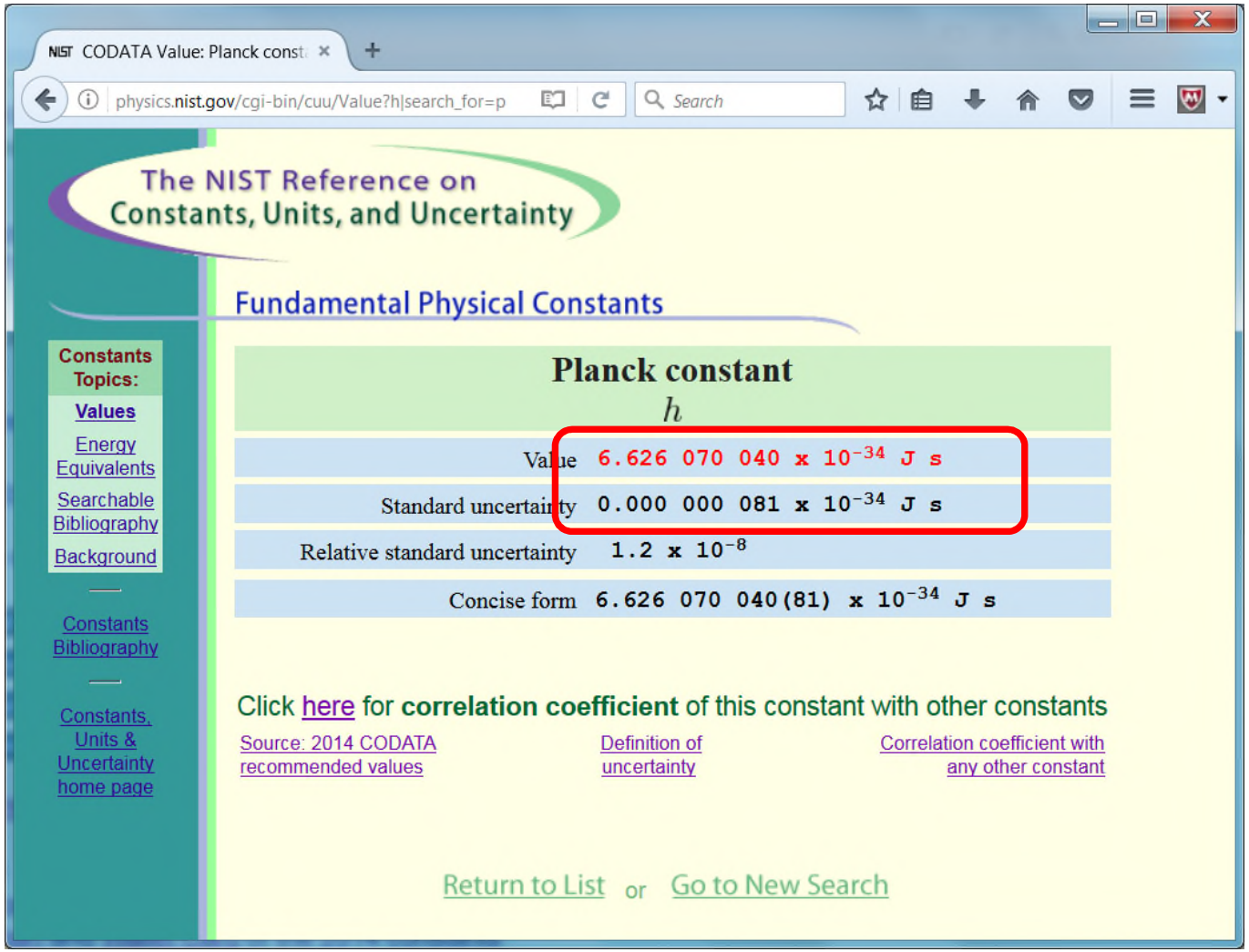


Overload of metadata such as the *method* should be avoided.

Better is to have a simple quantity and capture method in as an accuracy.

This is what phsyics does.

Standards: ***accuracy***

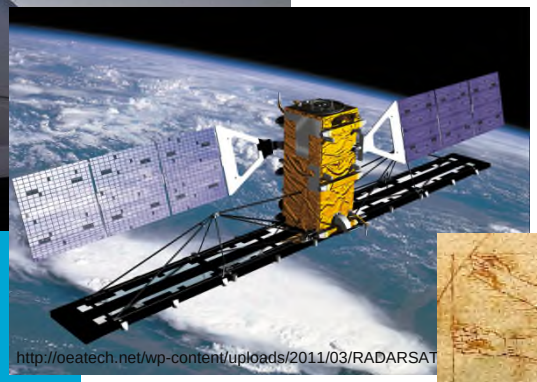


TU Delft

Van Oord
Marine ingenuity

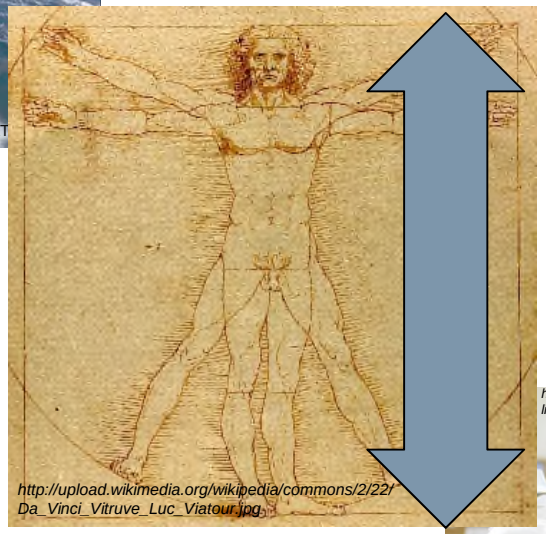
Community: as large as possible

Marine/coastal scales



SPACE

6400 km:
North Atlantic
Oscillation (NAO)



gap 10^{-6} > 10^{+6}

Water scarcity
Flood protection
Water pollution

CERN

suspended
mud particle
64 µm

TU Delft

Van Oord
Marine ingenuity

What is OpenEarth

2008

- Make **existing** data, models and tools available
  - in crease efficieny of projects
  - prevent double work
  - prevent loss of previous work
    - > due to lack of archiving (no time to store)
    - > due to new bugs (no time to test)
  - make work nicer: less maintenance, more development

- **a data and source code repository**
- **a community**

  a repository is useless without people using it
- **a philosophy**

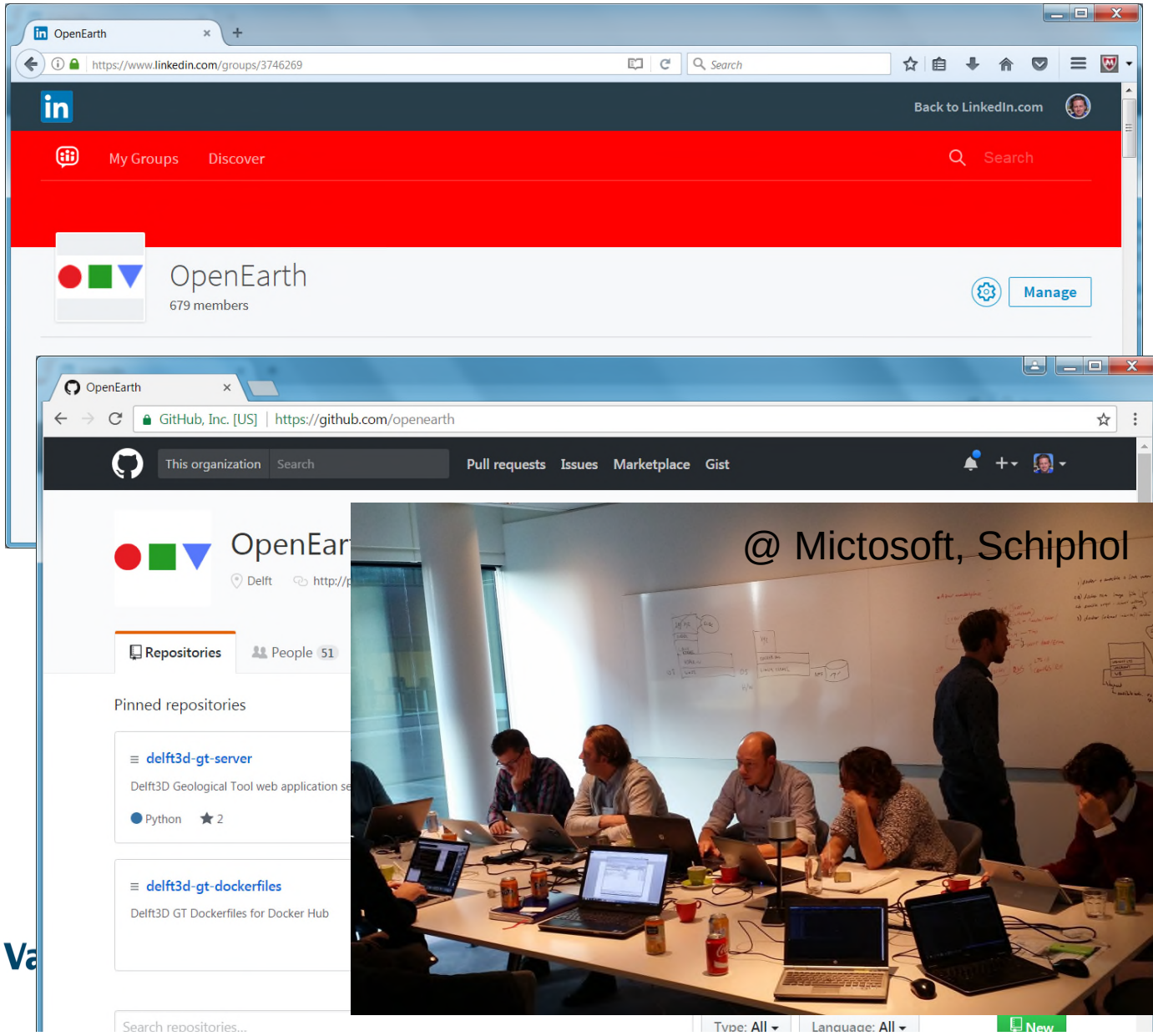
  a community is useless without collaboration: **cooperate!**

Deltare

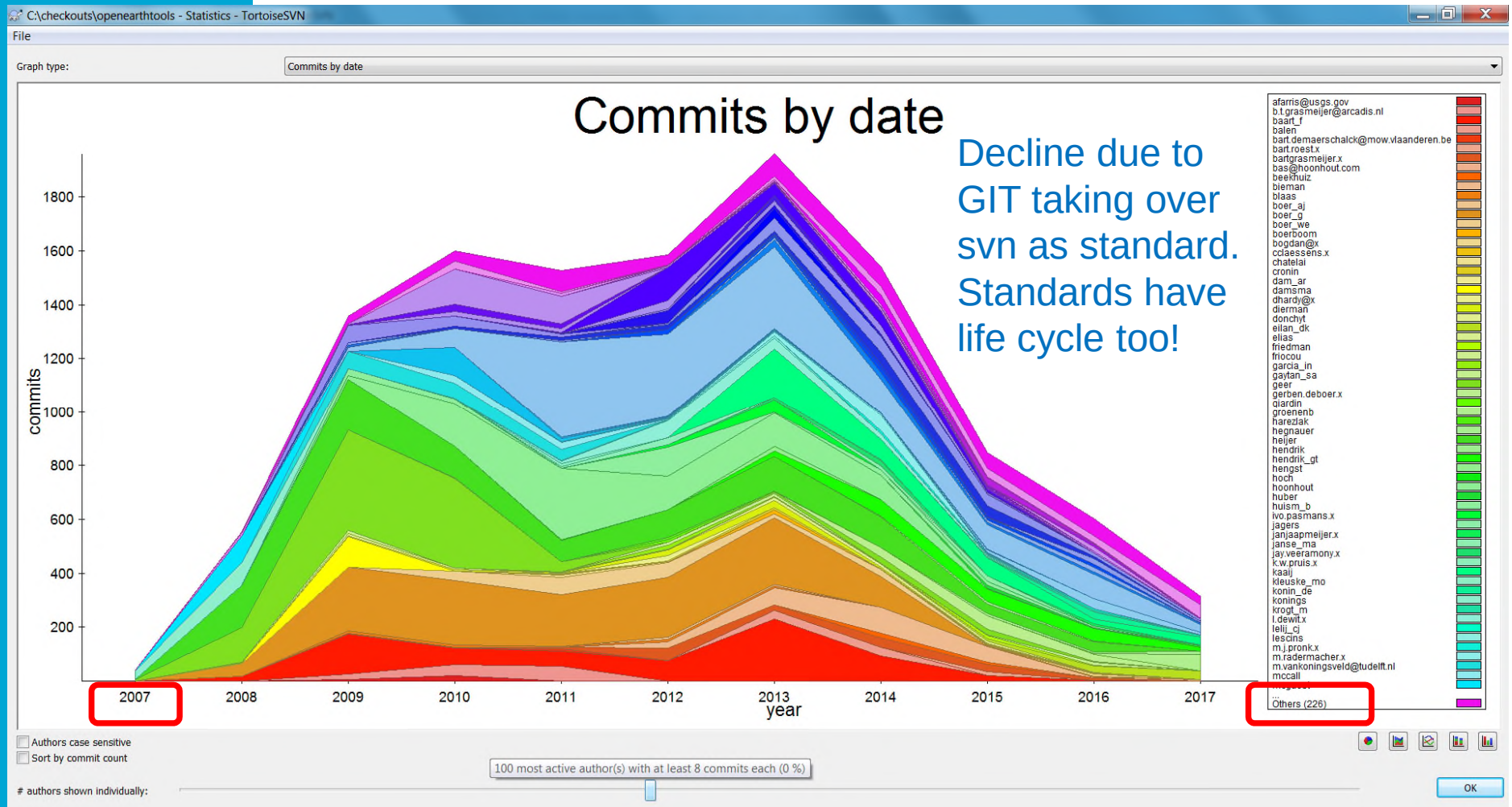
Community: social events/training

2 former PhD students just started OpenEarth by combining their finished work: **Mark van Koningsveld** (UT: coastal sand) & **Gerben de Boer** (TUD: marine mud)

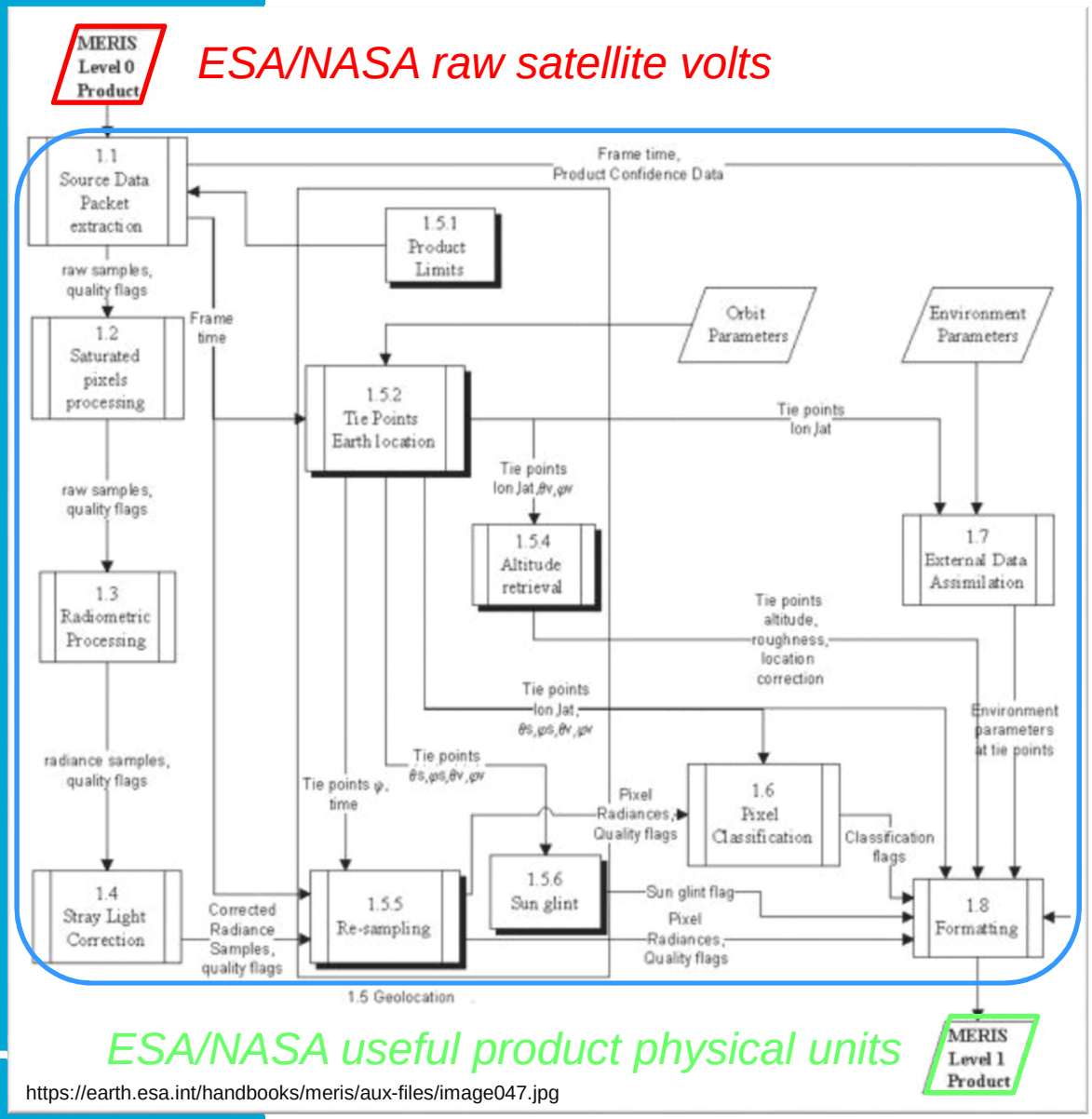
Social network has grown to **600+ users**, incl 3rd world + Fortune 500.



Community: global workplace



Philosophy: data = raw data + **tools**



https://earth.esa.int/handbooks/meris/aux-files/image047.jpg

data product =

raw data: volts counts → *+ software*

http://logfurniturehowto.com/wp-content/uploads/2012/07/Types-Of-Screwdrivers-And-Their-Uses.jpg

Philosophy: data = raw data + **tools**

raw *data*

scripts

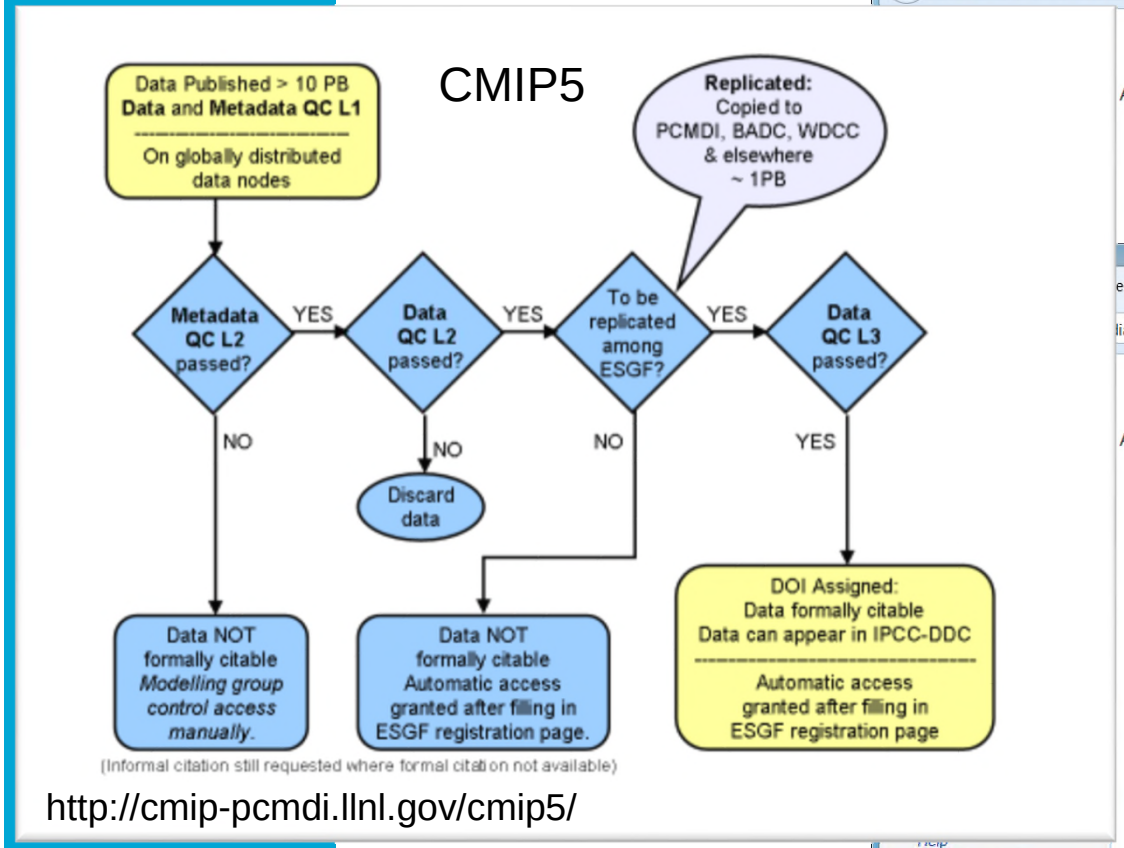
under version control

tailored *data*

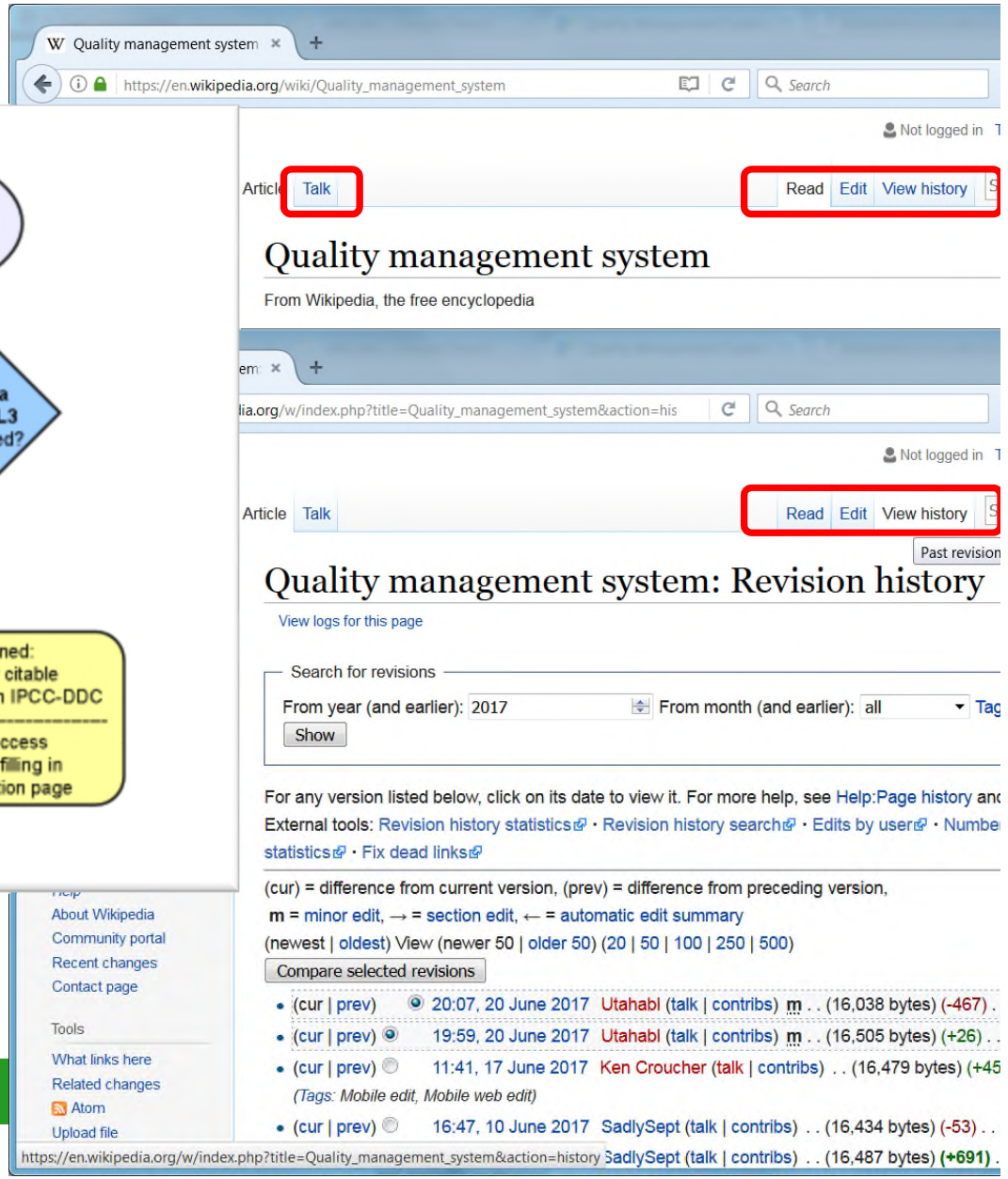
TU Delft

Van Oord
Marine ingenuity

Philosophy: good process > good result



http://cmip-pcmdi.llnl.gov/cmip5/



TU Delft

Van Oord
Marine ingenuity

Philosophy: good process > good result

Any company needs audited ISO 9001 workflows as a license to operate.
Science gets away without being audited on workflows. This needs to change.

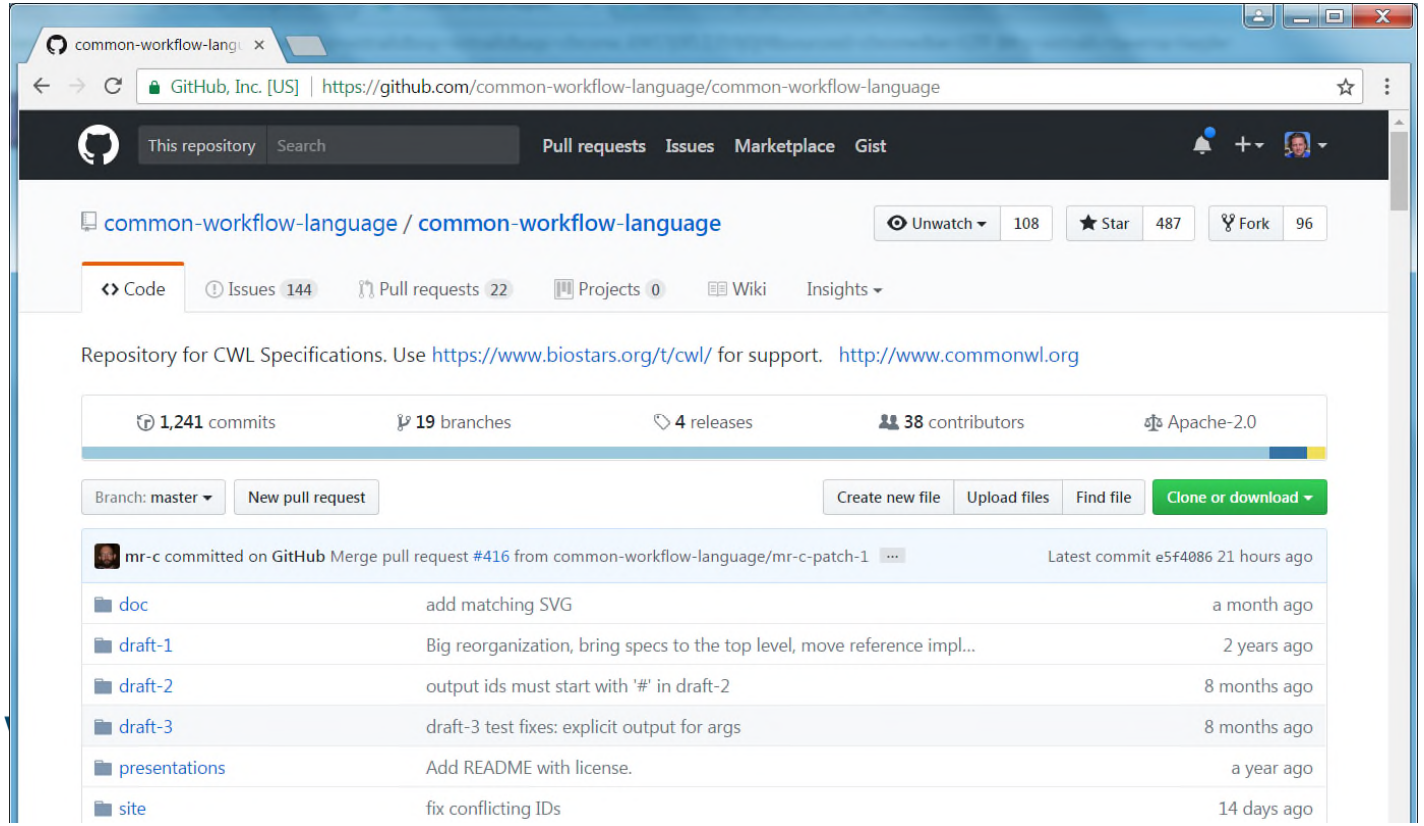
TU Delft

Van Oord
Marine ingenuity

Philosophy: tool version control

Graphical workflow tools instead of scripts

- VisTrails https://www.vistrails.org
- Taverna http://www.taverna.org.uk
- Kepler https://kepler-project.org
- CWL: https://github.com/common-workflow-language



Philosophy: tool version control

http://www.knups.nl/leuk_plaatje/56/Helaas_pindakaas.html

Are you Peanut butter Principle Proof? (PPP)

Philosophy: tool testing

We scientist need to work on sharing our analysis tools for reproducibility, lineage etc.

Use standard unit test for your scripts.
Before you apply them to your actual raw data

- Mass conservation
- Positive definite (concentrations)
- …

http://www.nature.com/news/2010/101013/full/467775a.html

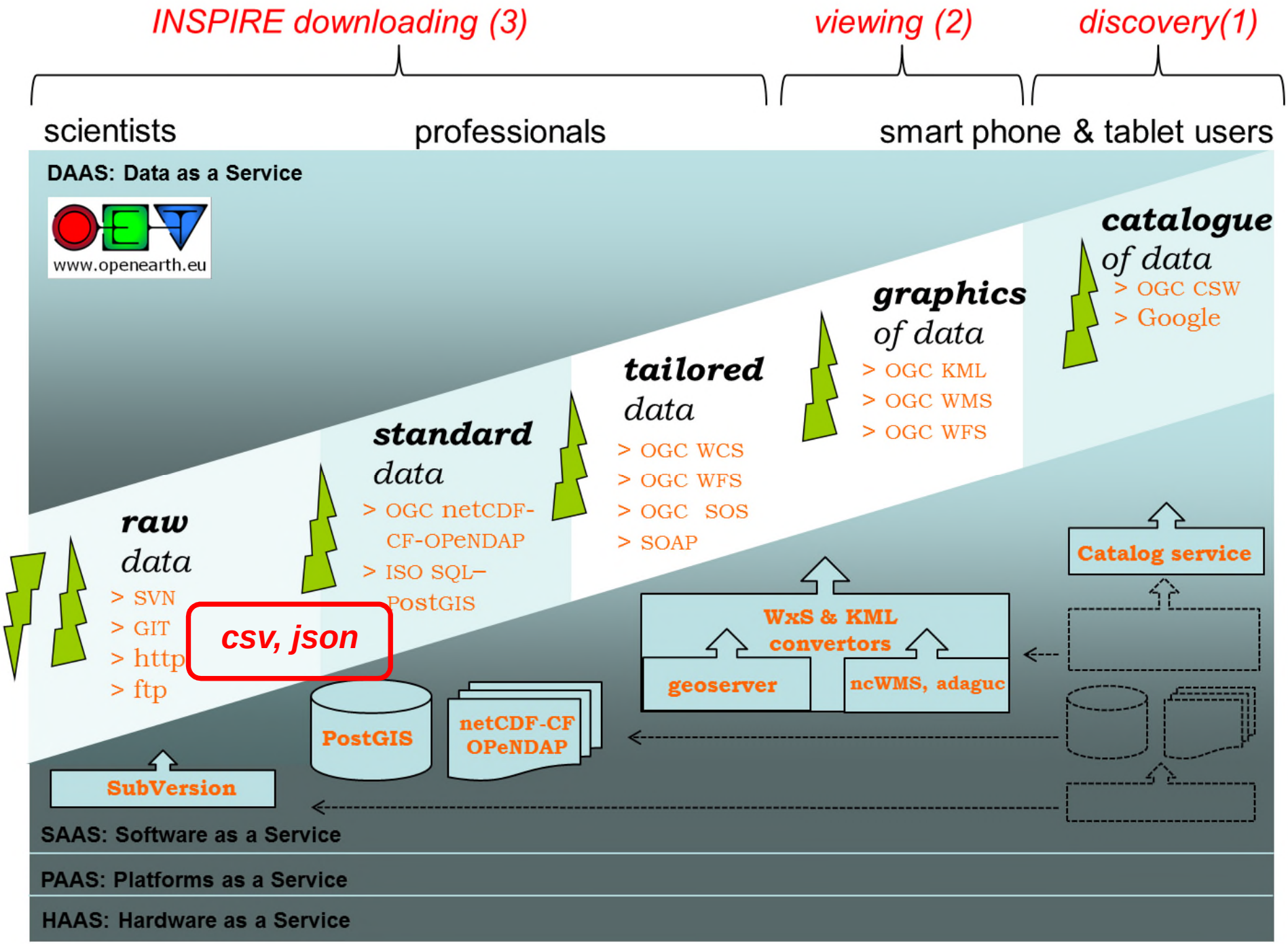
Philosophy: ETL > ELT

Classic Business scripts: ETL

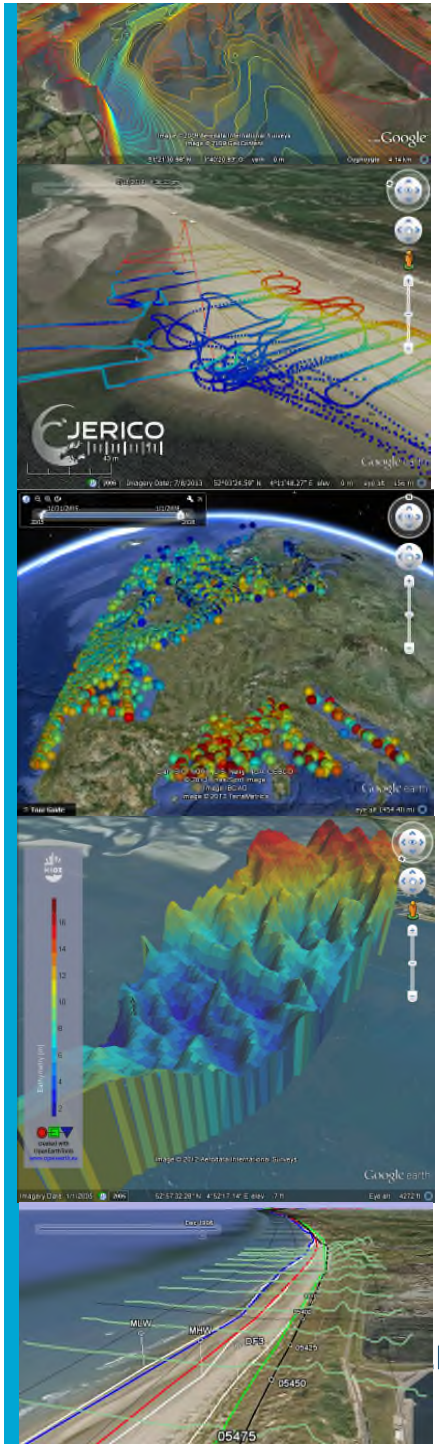
- Extract: raw data
- Transform: to standard data (SQL, netCDF)
- Load: these standard data

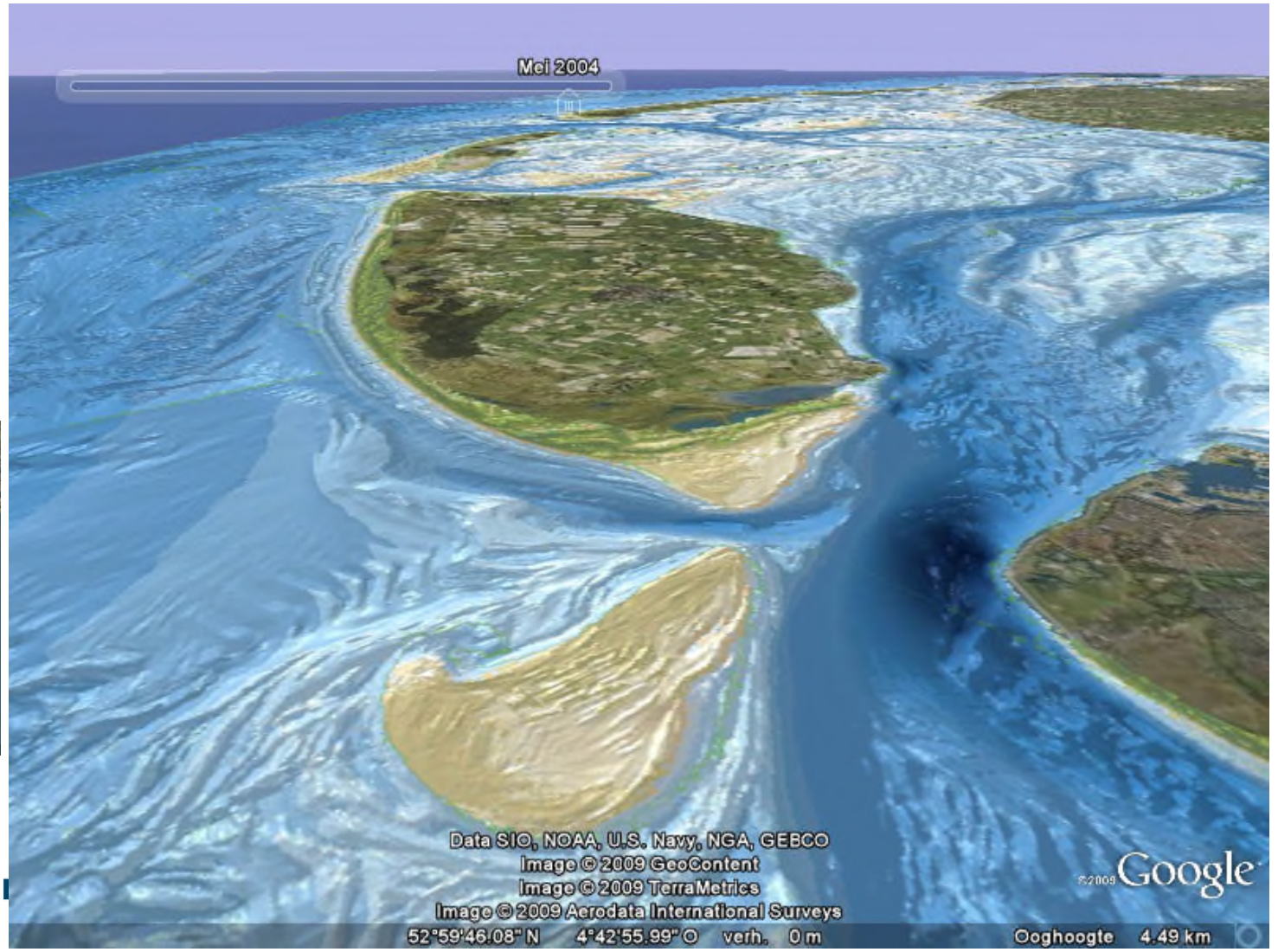
In Big Data, Business are moving to noSQL Databases.
This implies a swap of T and L, and skipping of standards.

- Extract: raw data
- Load: the ***raw*** data (***csv, json***)
- Transform: load the raw data on-the-fly to in-memory structures (Spark, Pandas DataFrame, …)


INSPIRE downloading (3)
viewing (2)
discovery(1)
scientists
professionals
smart phone & tablet users
DAAS: Data as a Service
www.openearth.eu
raw
data
> SVN
> GIT
> http
> ftp
standard
data
> OGC netCDF-CF-OPeNDAP
> ISO SQL–PostGIS
tailored
data
> OGC WCS
> OGC WFS
> OGC SOS
> SOAP
graphics
of data
> OGC KML
> OGC WMS
> OGC WFS
catalogue
of data
> OGC CSW
> Google
csv, json
PostGIS
netCDF-CF OPeNDAP
SubVersion
WxS & KML convertors
geoserver
ncWMS, adaguc
Catalog service
SAAS: Software as a Service
PAAS: Platforms as a Service
HAAS: Hardware as a Service

Powerful graphics needed to sell standards: e.g. Google Earth kml





Adopted by companies too

OpenEarth: A Knowledge Management Workflow for Dredging Projects 3

M. VAN KONINGSVELD, T. DAMSMA, R. VAN DER HOUT, J. VAN WIECHEN AND G. DE BOER

OPENEARTH: A KNOWLEDGE MANAGEMENT WORKFLOW FOR DREDGING PROJECTS

ABSTRACT

Research and consultancy as well as construction projects often spend a significant part of their budget to set up some basic infrastructure for data and knowledge management, most of which dissipates again once the project is finished. Standing initiatives so far have not been successful in providing a proper data and knowledge management system for data, models and tools. OpenEarth (www.openearth.eu) was developed as a free and open source alternative to the current often ad-hoc approaches to deal with data, models and tools.

OpenEarth as a whole (philosophy, user community, infrastructure and workflow) is the first comprehensive approach to handling data, models and tools that actually works in hydraulic engineering practice at a truly significant scale. It is implemented effectively not only at its original founding organisations, Delft University of Technology and Deltares, but also in a number of sizeable research programmes with multiple partners (such as research programme "Building with Nature" with 19 partners from one country) and from multiple countries (such as the 3-year European Union FP7 research programme MICORE with 15 partners from 9 countries). It has been adopted as the main data management workflow for all research programmes around the Sand Engine Delfland and was awarded the Dutch Data Prize 2012 for technical sciences by 3TU.datacentrum, the data archiving institute of the Dutch technical universities, and DANS, the data archiving institute of the Dutch National Science Foundation (NWO) and the Royal Dutch Academy of Sciences (KNAW).

For data, models and tools that are truly strategic and really cannot be shared, OpenEarth stimulates the set-up of internal OpenEarth clones. This way the OpenEarth workflow can still be adopted, promoting collaboration within an organisation, while taking care of security considerations at the same time.

This article is based on and updates the OpenEarth philosophy, infrastructure and main workflow protocols as presented at WODCON XIX in Beijing, China (Van Koningsveld *et al.*, 2010). A number of practical example applications that have been realised to date are given to illustrate OpenEarth's potential for the dredging industry.

Above: OpenEarth (www.openearth.eu) was developed as a free and open source alternative to the current often ad-hoc approaches to deal with data, models and tools, adopting two existing web services that are fully operational with a large community of users, **OPeNDAP** protocol for accessing data numbers and **Google Earth KML** standard for accessing data graphics.

INTRODUCTION

The sustainable interaction between humankind and planet Earth poses huge hydraulic and environmental engineering challenges. Confronting these challenges one-project-at-a-time, while seemingly attractive from a budget management perspective, results in grave inefficiencies in developing and archiving the basic elements that are invariably involved: data, models and tools. Hardly any project is by itself of sufficient scale to develop easily accessible and high-quality data archives, state-of-the-art modelling systems and well-tested analysis tools under version control. Research, consultancy as well as major construction projects commonly spend a significant part of their budgets to set up some basic data and knowledge management infrastructure, most of which dissipates again once the project is finished.

Internally institutions generally employ intranet services and internal networks to collaborate and exchange information. However, owing to increasing complexity, large projects nowadays are regularly executed by consortia.

Van Oord

Marine ingenuity

Van Oord has a HQ department under Engineering & Estimating dedicated to OpenEarth Data Management.

Van Oord is a marine contractor for land reclamation, dredging, oil & gas and offshore wind.

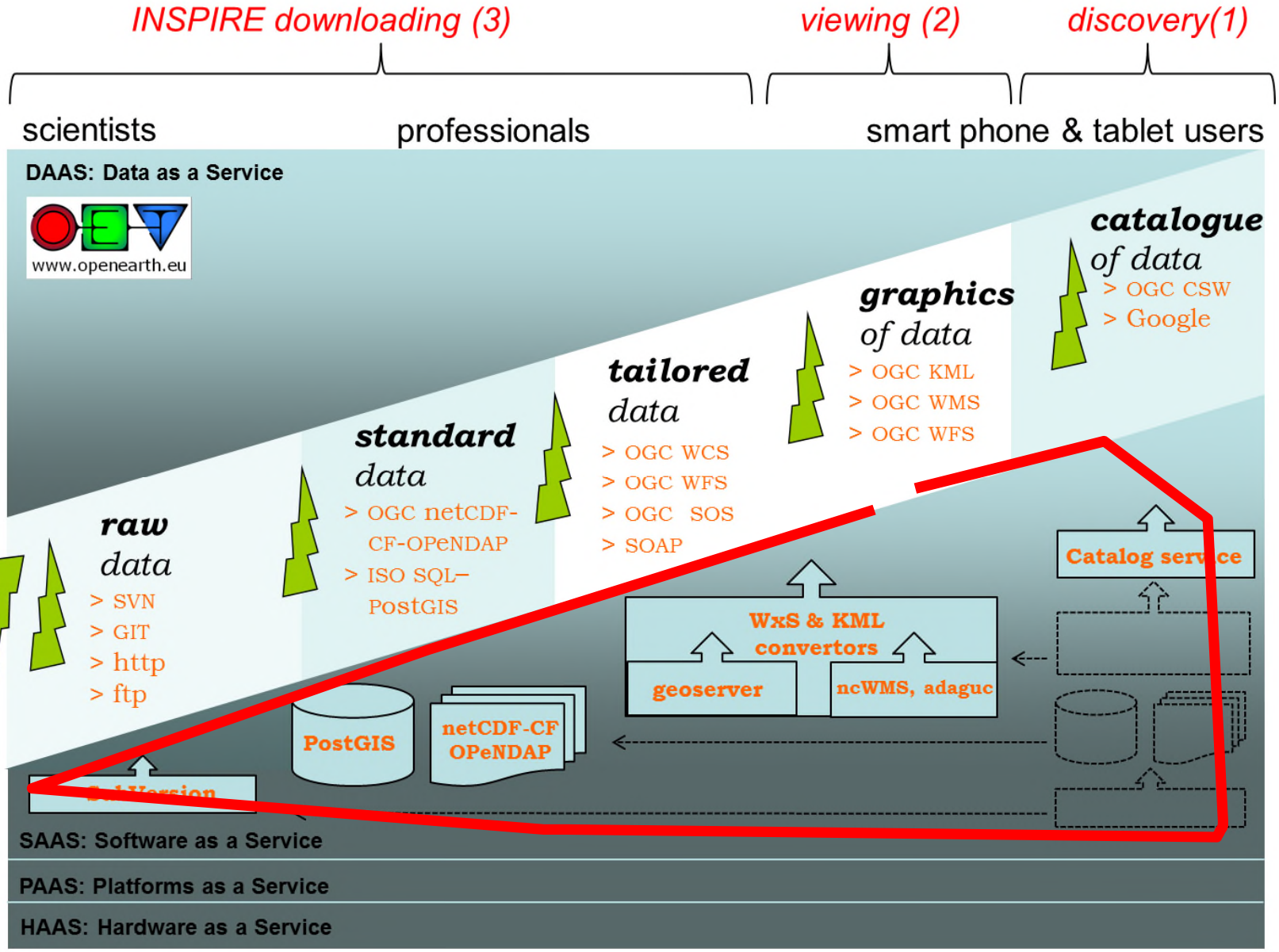
https://www.vanoord.com

https://www.iadc-dredging.com/ul/cms/terraetaqua/document/3/7/3/373/373/1/article-openeartha-knowledge-management-workflow-for-dredging-projects-terra-et-aqua-131-1.pdf

TUDelft

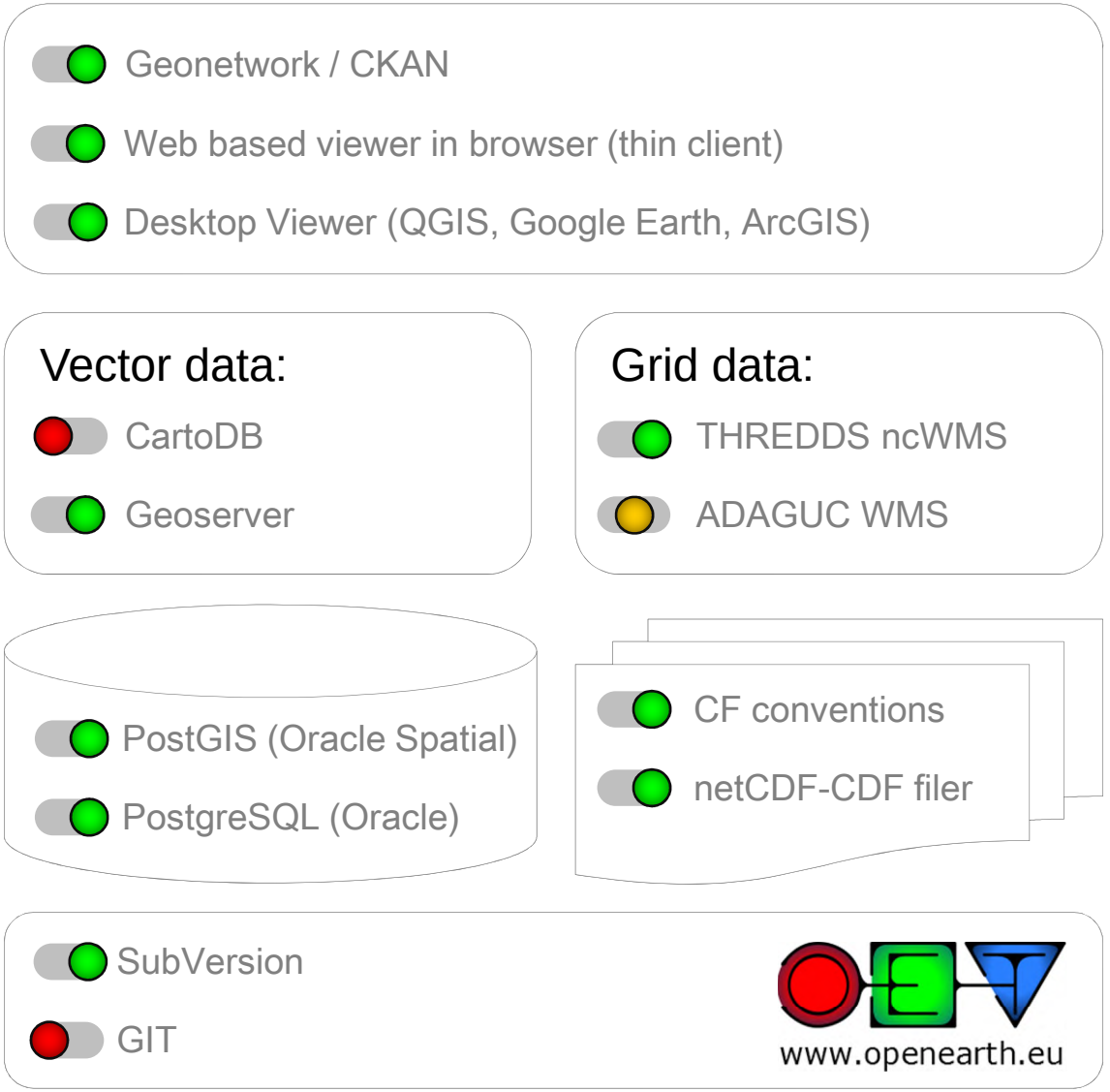
Van Oord Marine ingenuity

Webservices need a server stack



TU Delft

Server stack has components



catalogue *of data*

graphics *of data*

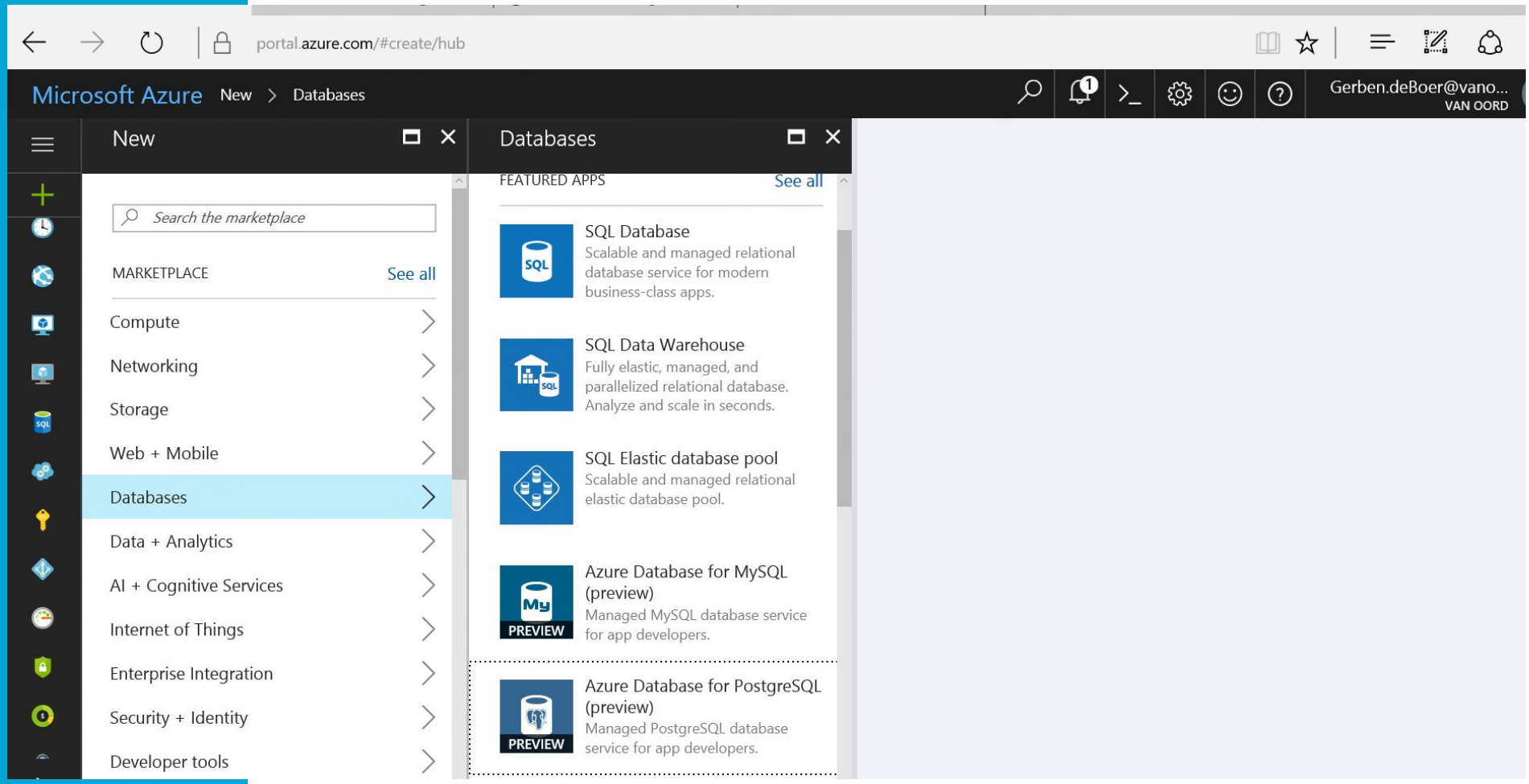
tailored *data*

standard *data*

raw *data*

TU Delft

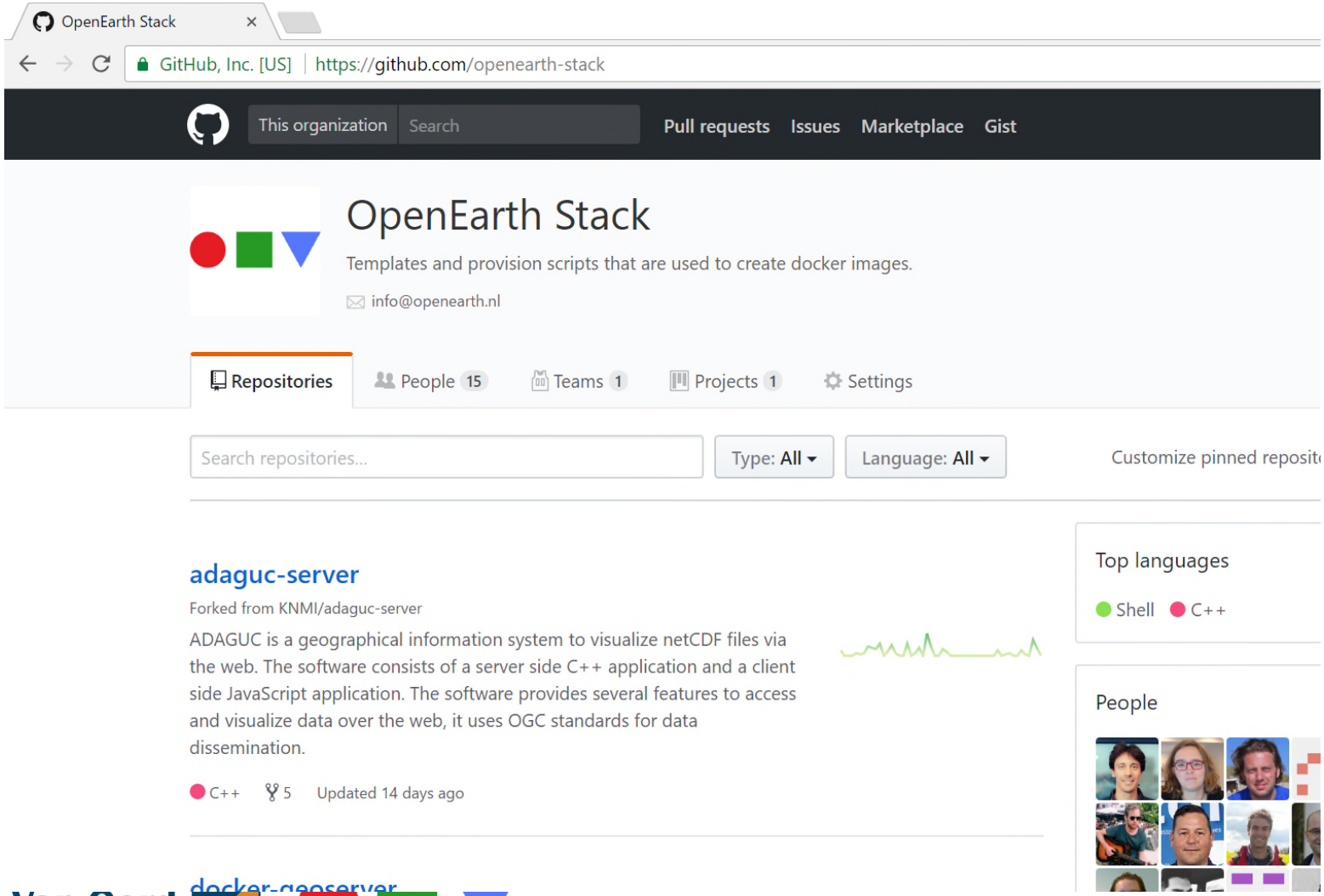
Server stack can be hosted in cloud



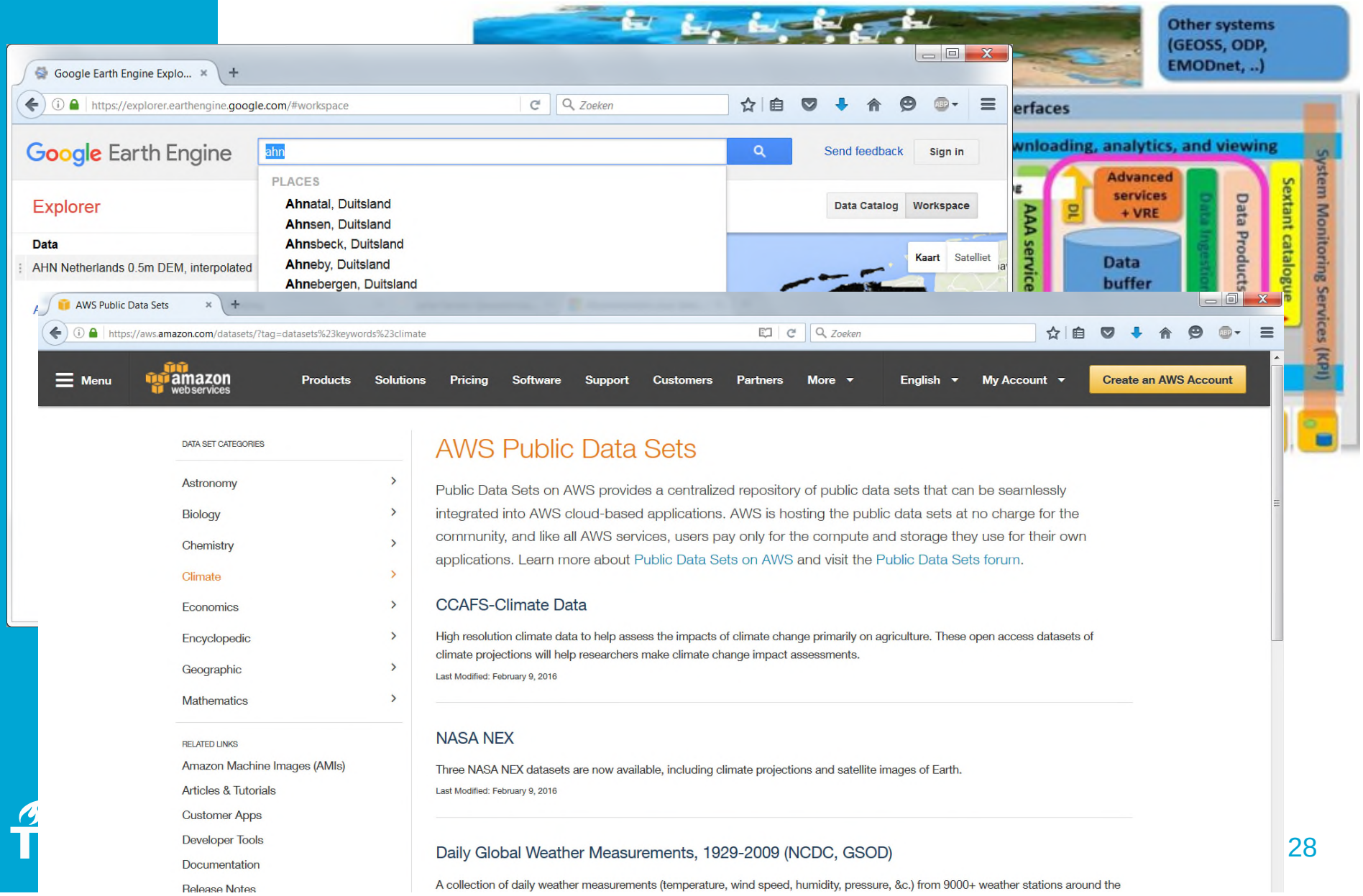
TU Delft

Van Oord
Marine ingenuity

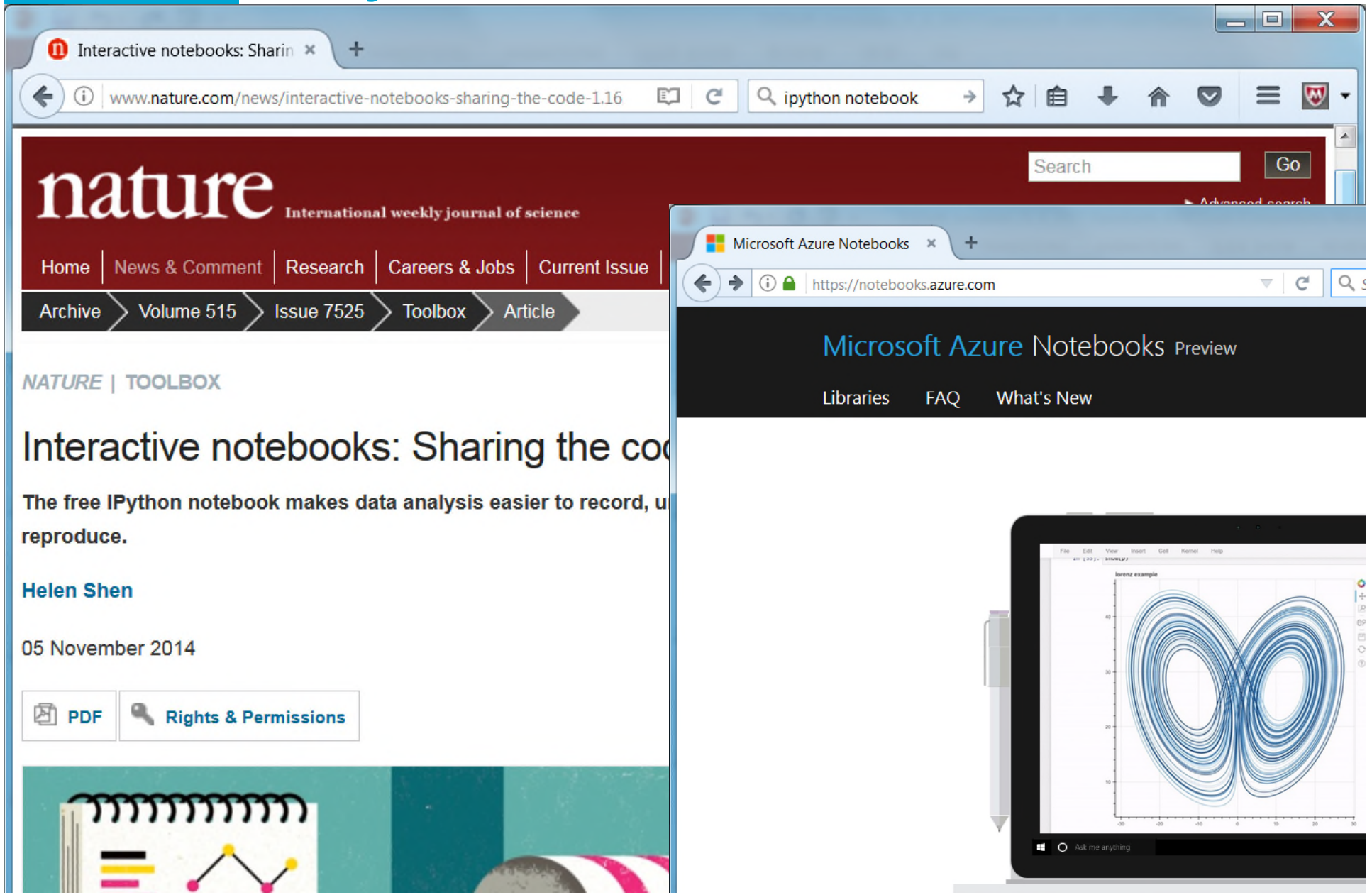
Server stack can be hosted in cloud



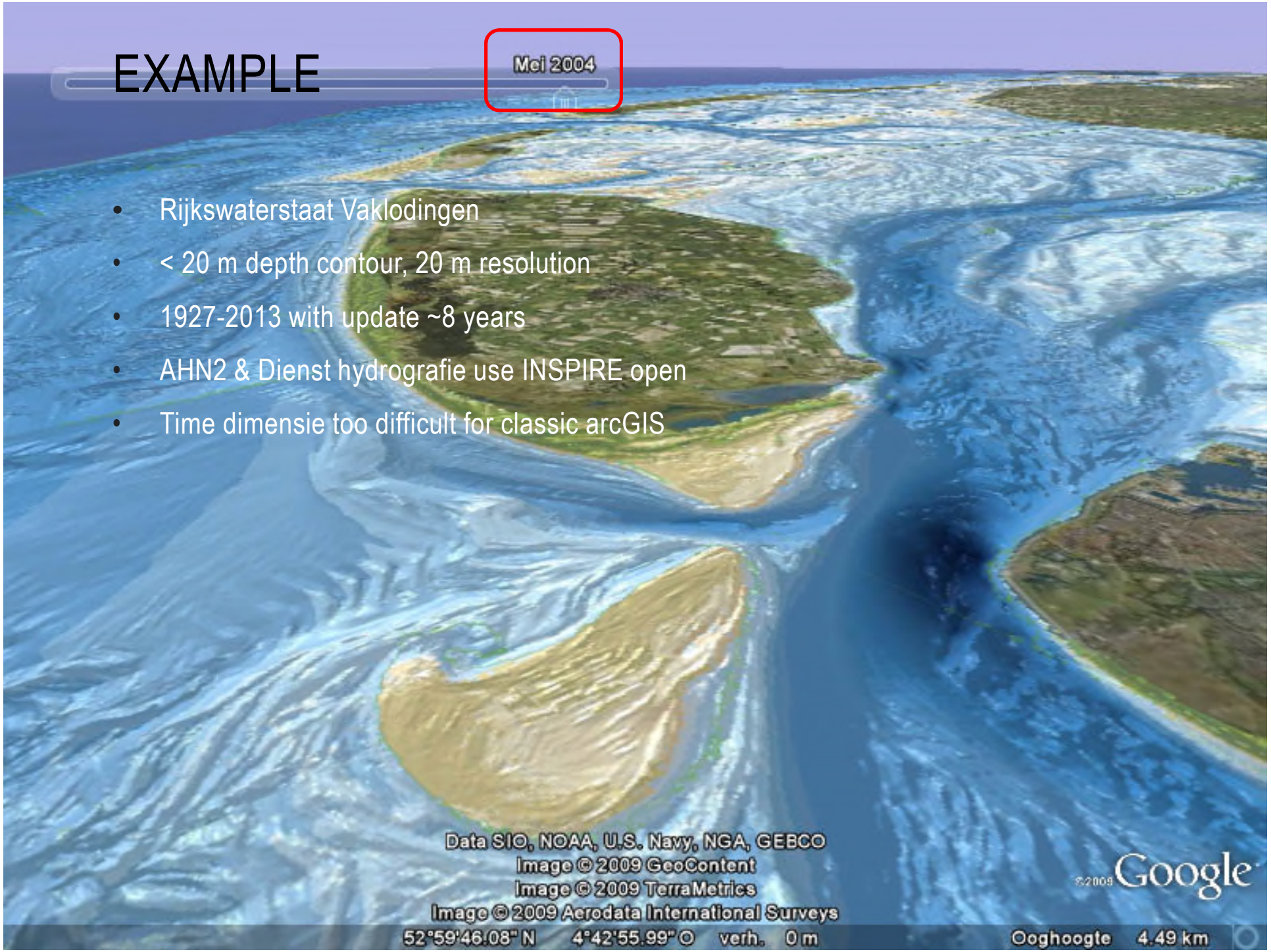
Beyond webservices: Datalakes



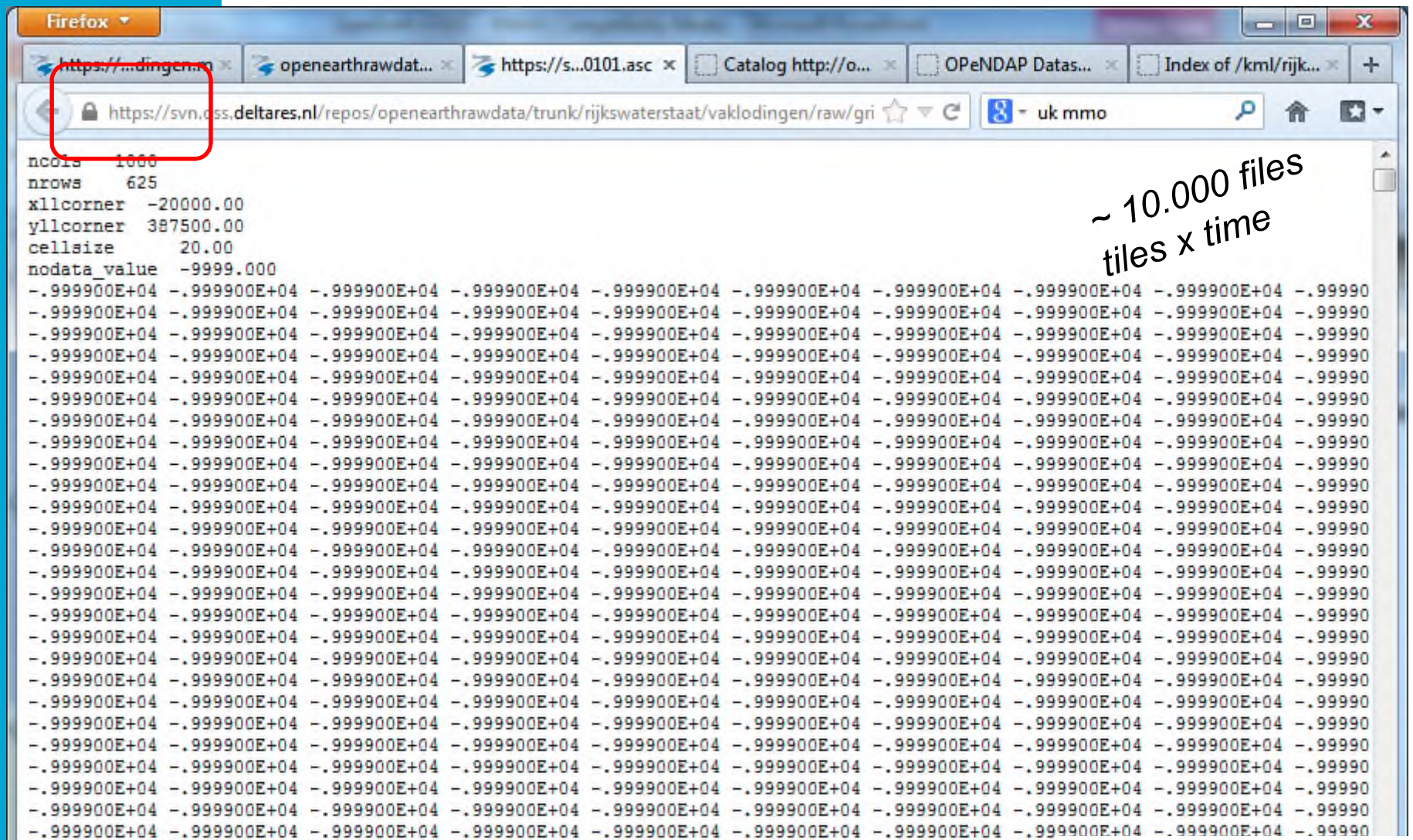
Beyond webservices: Notebooks



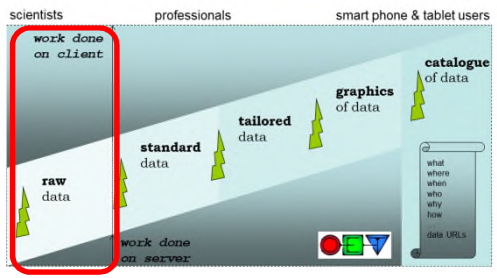
EXAMPLE



- Rijkswaterstaat Vaklodingen
- < 20 m depth contour, 20 m resolution
- 1927-2013 with update ~8 years
- AHN2 & Dienst hydrografie use INSPIRE open
- Time dimensie too difficult for classic arcGIS



Marine ingenuity



Firefox

https://...dingen.m | openearthrawdat... | https://s...0101.asc | Catalog http://o... | OPeNDAP Datas... | Index of /kml/rijk...

opendap.deltares.nl/thredds/dodsC/opendap/rijkswaterstaat/vaklodingen/vaklodingenKB140_

uk mmo

OPeNDAP Dataset Access Form

~200 files
tiles = f(time)

Tested on Netscape 4.61 and Internet Explorer 5.00.

Action: Get ASCII | Get Binary | Show Help

Data URL: http://opendap.deltares.nl/thredds/dodsC/opendap/rijkswaterstaat/vakloc

Global Attributes:

```
Metadata_Conventions: CF-1.5
title: vaklodingen
summary: bathymetry and topography measurements along the Dutch coast
keywords: bathymetry, coast
keywords_vocabulary: http://www.eionet.europa.eu/gemet
standard_name_vocabulary: http://cf-pcmdi.llnl.gov/documents/cf-standard-names/
history: Created on 2012-09-11 16:11:41 by Kees den Heijer on computer
DELTARES\D00553 with script $Id: ncgen vaklodingen.m 5343 2012-09-11 14:11:01Z
```

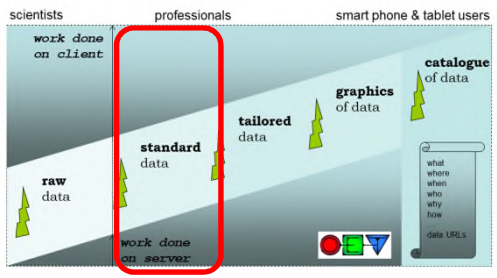
Variables: x: Array of 32 bit Integers [x = 0..499]

x:

```
standard_name: projection_x_coordinate
long_name: x-coordinate
units: m
    ition:
    _mapping: crs
    l_range: 290010.0, 299990.0
    lution: 20.0
```

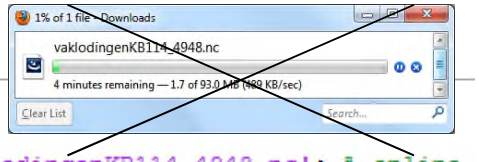
CERTIFIED OGC® COMPLIANT

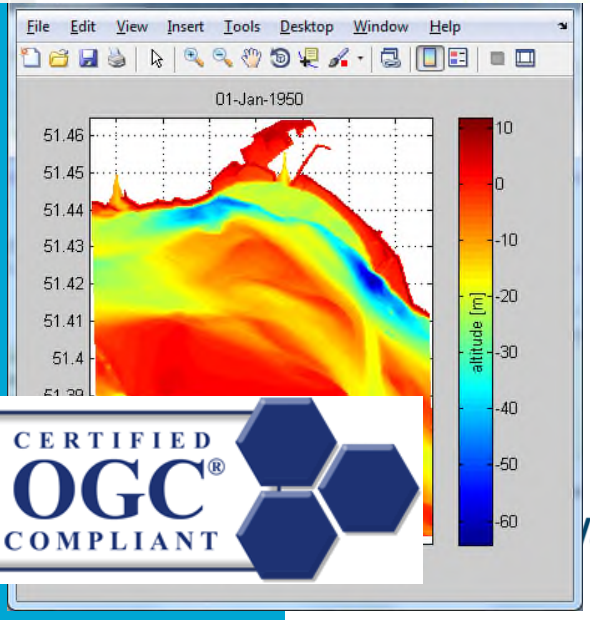
Marine ingenuity

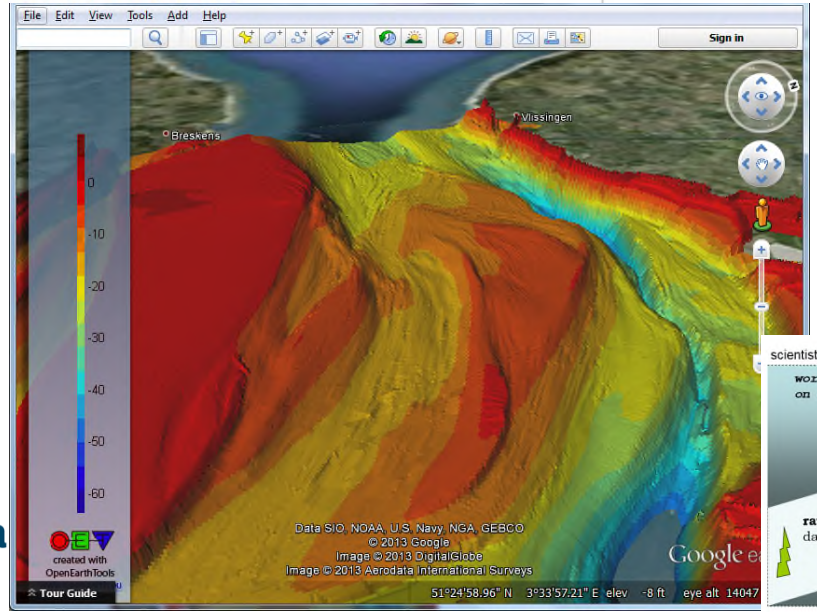


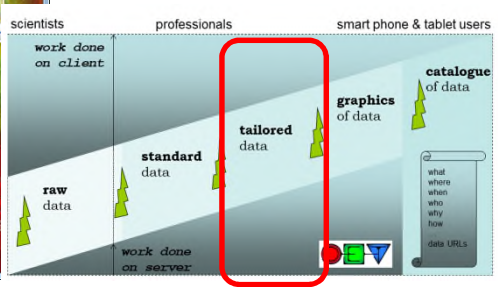
HOW SCIENTISTS USE DATA

```
% oetsettings
%% load web-data
url = 'D:/download/vaklodingenKB114_4948.nc'; % offline cache
url = 'http://opendap.deltares.nl/thredds/dodsC/opendap/rijkswaterstaat/vaklodingen/vaklodingenKB114_4948.nc'; % online
D.x           = ncread(url,'x');
D.y           = ncread(url,'y');
[D.x,D.y]     = meshgrid(D.x,D.y);
[D.lon,D.lat] = convertCoordinates(D.x,D.y,'CS1.code',28992,'CS2.code',4326); % RD to WGS84
D.time        = nc_cf_time(url); % 38 times
D.z           =  ncread(url,'z',[1 1 length(D.time)],[Inf Inf 1])'; % get last z
D.label       = [nc_attget(url,'z','long_name'),' [',nc_attget(url,'z','units'),']']
%% plot data
pcolorcorcen(D.lon,D.lat,D.z);
title(datestr(D.time(1)))
colorbarwithvtext(D.label)
axislat; grid on; tickmap('ll','fmt','%g')
%% plot in Google Earth
KMLsurf(D.lat,D.lon,D.z, 'fileName',[vaklodingenKB114_4948,'_',datestr(D.time(end)),'.kmz'],...
                         'zScaleFun',@(z)(z+50).*5);
```







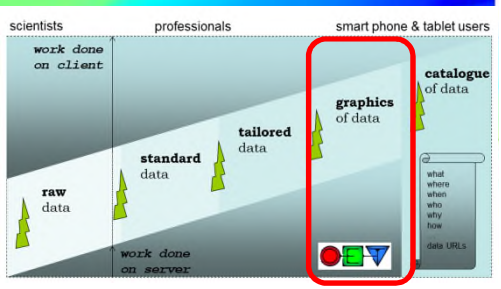


CERTIFIED OGC® COMPLIANT

WMS + TIJD + NASA COLORSCALE EXTENSION

http://opendap.deltares.nl/thredds/wms/opendap/rijkswaterstaat/vaklodingen/vaklodingenKB121_2120.nc?
service=WMS
&version=1.3.0
&request=GetMap
&bbox=4.5672207,52.9343465,4.7177535,53.0474114
&layers=z
&format=image/png
&crs=EPSG%3A4326
&width=800
&height=600&
styles=boxfill/ferret
&TIME=2012-01-01
&COLORSCALERANGE=-50,50

CERTIFIED OGC® COMPLIANT



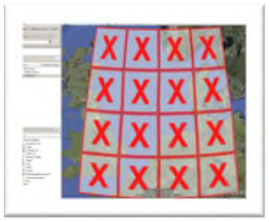
Firefox

Index of /kml/rijkswaterstaat/vaklodinge...

kml.deltares.nl/kml/rijkswaterstaat/vaklodingen/2012-01-01/

Google

Index of /kml/rijkswaterstaat/vaklodingen/2012-01-01

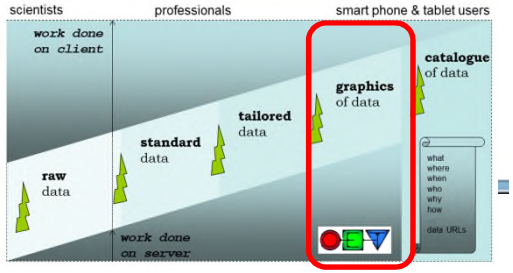


- Parent Directory
- 2012-01-01_0.kml
- 2012-01-01_0.png
- 2012-01-01_01.kml
- 2012-01-01_01.png
- 2012-01-01_012.kml
- 2012-01-01_012.png
- 2012-01-01_0120.kml
- 2012-01-01_0120.png
- 2012-01-01_01202.kml
- 2012-01-01_01202.png
- 2012-01-01_012022.kml
- 2012-01-01_012022.png
- 2012-01-01_0120221.kml
- 2012-01-01_0120221.png
- 2012-01-01_01202210.kml
- 2012-01-01_01202210.png
- 2012-01-01_012022102.kml
- 2012-01-01_012022102.png
- 2012-01-01_0120221022.kml
- 2012-01-01_0120221022.png
- 2012-01-01_01202210222.kml
- 2012-01-01_01202210222.png

klodingen/2012-01-01/2012-01-01_0120221022.kml

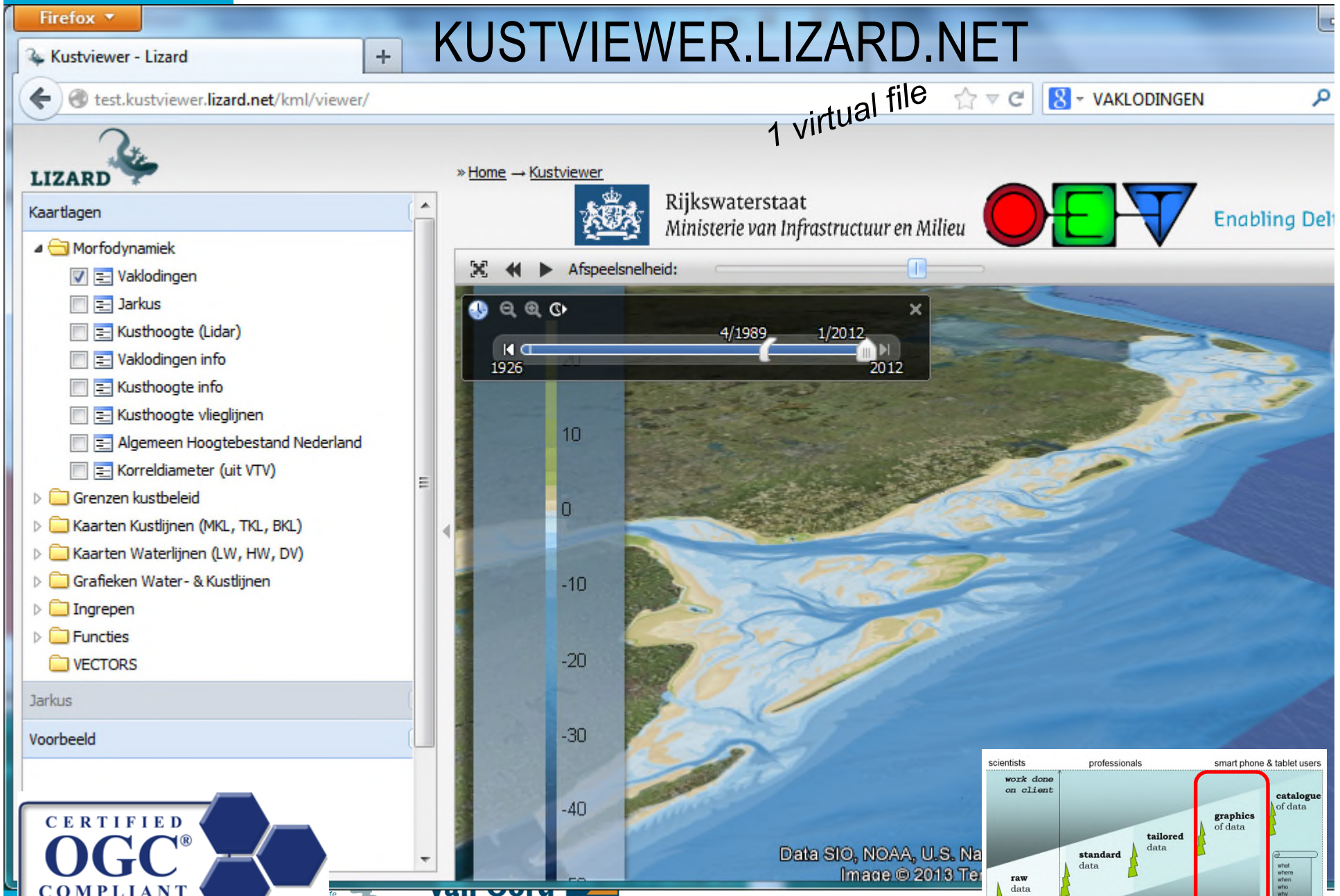
CERTIFIED OGC® COMPLIANT

Marine ingenuity



KUSTVIEWER.LIZARD.NET

1 virtual file



CERTIFIED OGC® COMPLIANT

Marine ingenuity

Firefox

https:...ngen.m | openearthra... | https:/...01.asc | Catalog http:... | OPeNDAP Da... | Index of /kml... | vaklodingen ... | Dataset d...

https://publicwiki.deltares.nl/display/OET/Dataset+documentation+Vaklodingen

VAKLODINGEN

Dashboard › OpenEarth › ... › Dataset documentation
› Dataset documentation Vaklodingen

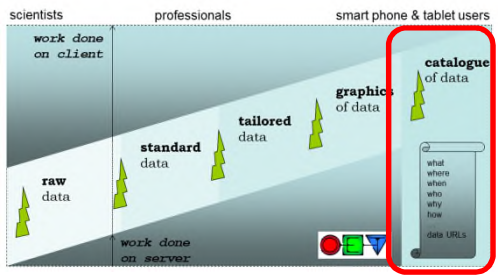
Browse Log In Sign Up Search

Dataset documentation Vaklodingen

Added by Gerben de Boer, last edited by Kees den Heijer on 08-11-2012 (view change)

Vaklodingen bathymetry data in OpenEarth:

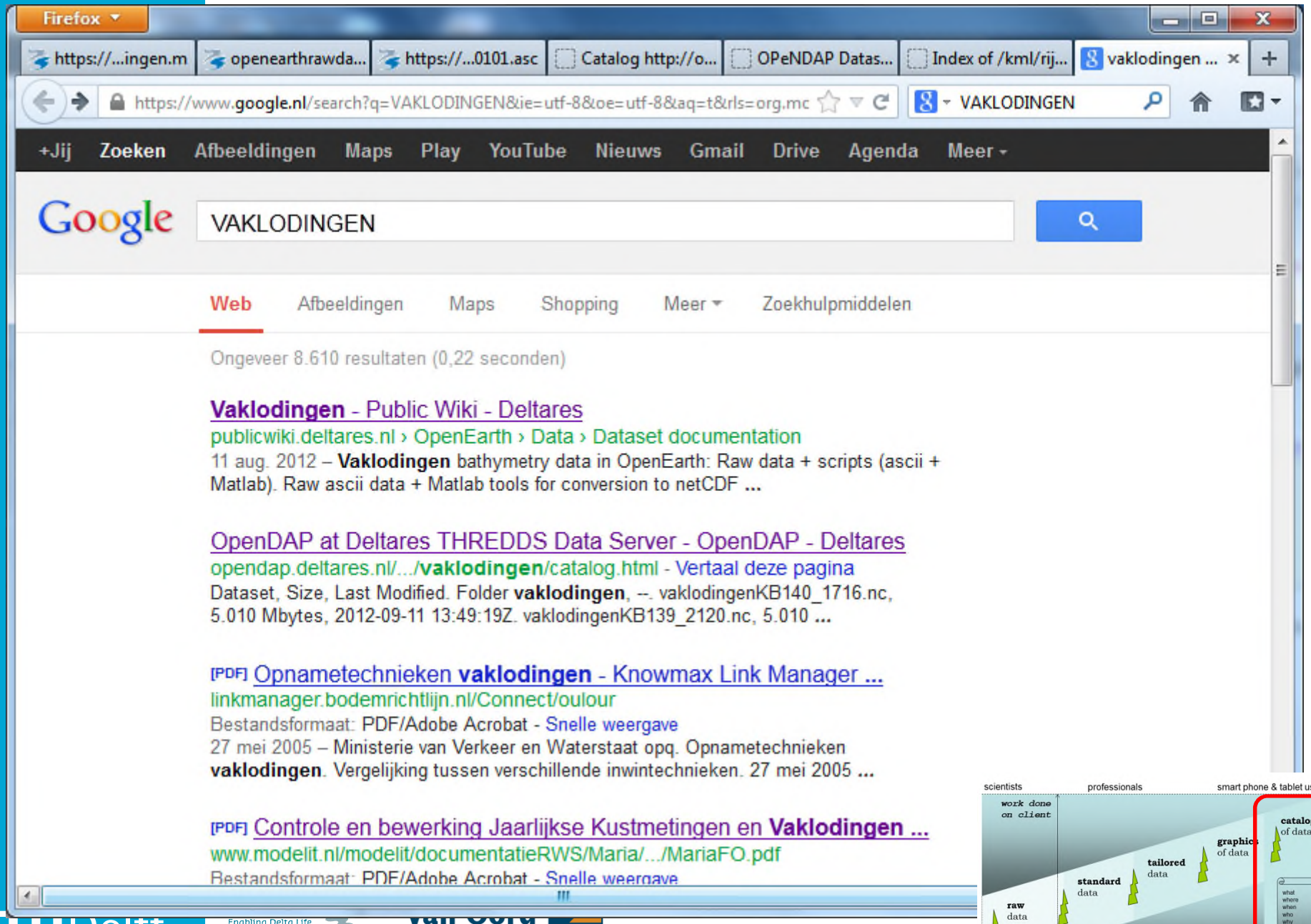
1. Raw data + scripts (ascii + Matlab)
   - Raw ascii data + Matlab tools for conversion to netCDF
     https://svn.oss.deltares.nl/repos/openearthrawdata/trunk/rijkswaterstaat/vaklodingen/ (register at oss.deltares.nl)
2. Standardized data (netCDF on OPeNDAP). The netCDF files can be viewed and analyzed interactively with UCIT.
   - 2012 release: production: http://opendap.deltares.nl/thredds/dodsC/opendap/rijkswaterstaat/vaklodingen/catalog.html, proc
     the new `nc_gen` in Matlab OpenEarthTools. This new release now has 80% (37/46 items) conformity with ISO meta-data sta
     with ncISO (we'll aim for more in release 2013 next year)
   - 2011 release: production: (processed with with the generic `nc_multibeam` netCDF tiles toolbox with `Rijkswaterstaat` vak
     tile configuration in Matlab OpenEarthTools). This old release has been deleted but can be regenerated any time with the o
     of the processing chain as kept in SubVersion.
   - 2010 release: production: (generated with abandoned dedicated processing line in Python). This old release has been dele
     can be regenerated any time with the old version of the processing chain as kept in SubVersion.
3. Visualized data (kml) (only 2012 release)
   - Vaklodingen overview (with links to OPeNDAP): http://kml.deltares.nl/kml/rijkswaterstaat/vak
   - All Vaklodingen data as tiled pngs, the same technology that Google Earth uses for all aeri
     /kml/rijkswaterstaat/vaklodingen.kml

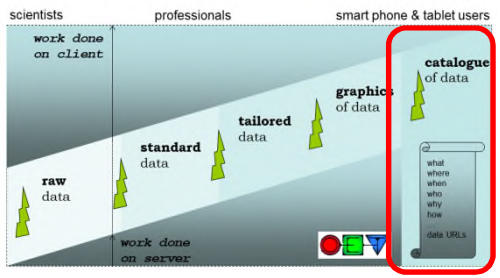


TU Delft

Enabling Delta Life

Van Oord
Marine ingenuity





TU Delft

Enabling Delta Life

Marine ingenuity

Recap

- There is a huge range of standards for 3 users types
  - Scientists > Professionals > Everyone, incl. general public
- Standard bodies are: ISO, OGC, INSPIRE, thematic groups
- Use existing de facto standards that are in general use already
- Avoid paper standard that have no reference implementation (QGIS)
- WGS84 for coordinates, avoid local systems (EPSG if you have to)
- Use UTC time coding + timezone
- Use standard lists for quantities and units: WoRMS, CAS, P01, CF
- Data = share data products and visuals
  - Raw data share also raw data
  - Scripts share your script via svn + git
- This makes guarantees lineage, tracebility etc
  - be Peanut butter Principcle Proof
  - NASA and ESA products work this way
  - Test your script before you apply them to raw data
- The cloud has entered science
  - Use central Data collections at Google, Amazon, Microsoft, EU
  - Use iPython notebooks to work with them
- SHARE YOUR WORK, or your work become irrelevant.

TU Delft

Van Oord
Marine ingenuity