Building the Canadian Spatial Data Foundry

An online portal for large scale spatial analysis

Pierre Racine
Research Assistant
Centre for Forest Research
Département des sciences du bois et de la forêt, Université Laval, Québec

Steve Cumming
Researcher
Centre for Forest Research
Département des sciences du bois et de la forêt, Université Laval, Québec

QUÉBEC 2012 GEOSPATIAL
CONGRÈS GÉOSPATIAL INTERNATIONAL
GLOBAL GEOSPATIAL CONFERENCE

Fondation canadienne pour l'innovation
Canada Foundation for Innovation

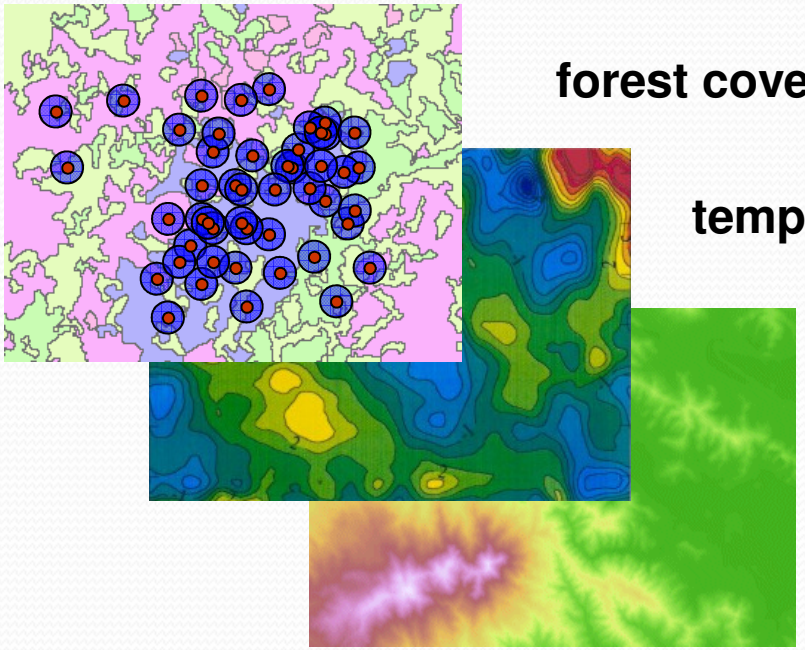
Boreal Avian Modelling Project

cfr Centre for Forest Research

UNIVERSITÉ LAVAL

Context I

- Researchers in forestry, ecology and environment
- Doing buffer analysis over HUGE raster and vector datasets (covering the extent of Canada)



forest cover

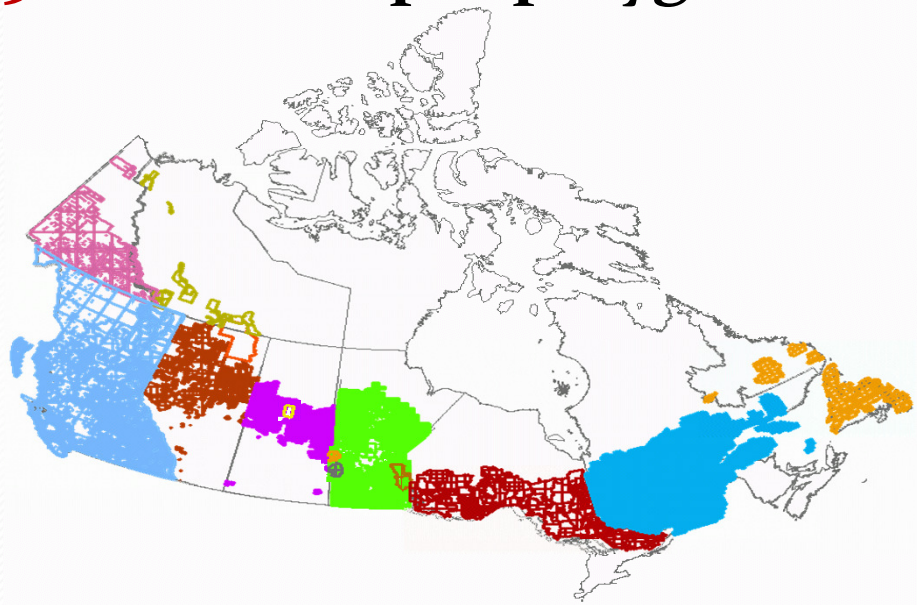
temperature

elevation, etc…

geom	obsID	cutProp	meanTemp	elevation	etc…
polygon	1	75.2	20.3	450.2	…
polygon	2	26.3	15.5	467.3	…
polygon	3	56.8	17.5	564.8	…
polygon	4	69.2	10.4	390.2	…
…		…	…	…	…

A Canada Wide Forest Inventory

- A Canadian wide forest inventory already exist but it is only based on a sampling
- A higher precision alternative is to standardize and merge 24 provincial photo interpreted inventories into a single one
- 26 000 000 polygons, 69 attributes per polygon
- 40 GB of vector data



Context II

Researchers must…

…learn lots of ArcGIS, to use few operations

…search for, download and assemble large datasets

- **historical data are often lost**
- **data are delivered in many different formats**
- **documentation is generally written for GIS experts**
- **datasets are too large to fit in one file (shp limited to 2 GB, complete forest cover for Canada is 40GB, complete DEM for Canada is 9 GB, etc…)**
- **computation is often too difficult for ArcGIS (800 buffers over 5 000 000 polygons)**

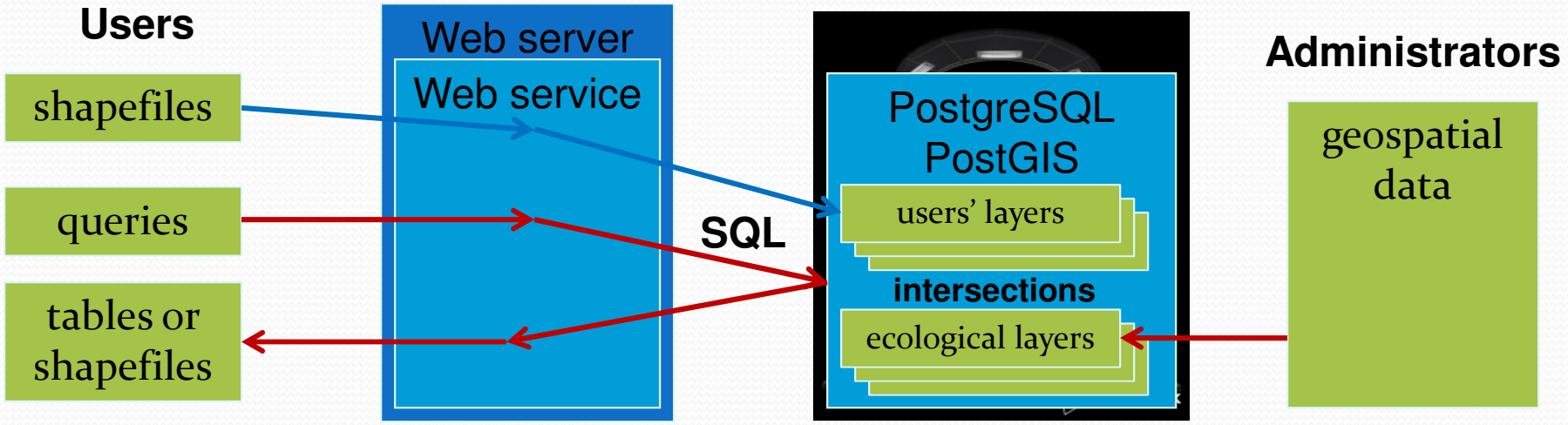
…struggle for weeks, if not months, to get their data table ready for statistical analysis…

In brief: researchers waste much energy on tasks unrelated to their main priority: research!

The Canadian Spatial Data Foundry

- A **web site** (or service)
- Backed by a **spatial database** hosted on a **supercomputer**
- **GIS administrators** upload **preassembled** datasets of **ecological layers** (vector & raster, historical data)
- **Users with accounts upload** their datasets (**shapefiles**)
- **Create intersection queries** with the ecological layers
- Obtain results in the form of **shapefiles** or **tables**

A Spatial Data Infrastructure (SDI) Dedicated to Large Scale Analysis



What Do We Need?

1. **A good storage solution**
   - HUGE datasets (26 000 000 polygons, 30 gig Landsat derived tiled raster coverages and many more)
2. **A robust and powerful raster/vector geoprocessing engine**
   - Intersecting SEVERAL, sometimes LARGE buffers with HUGE datasets
3. **A web interface to**
   - upload shapefiles in the store
   - document administrators and users datasets using a simple ISO 19115 profile
   - build intersection queries
6. **Optionally a web mapping engine**
   - To show datasets (and results) in a geographical context

Storage & Geoprocessing Options

- **Storage**
  - Files in the file system
  - Relational database (ArcSDE, Oracle, PostGIS)
- **Raster/vector geoprocessing engine**
  - Spatial database geoprocessing (ArcSDE, Oracle, PostGIS)
  - In house development (GEOS (C), JTS, SEXTANTE (Java), R)
- **Both solutions are intrinsically linked**
  - Storing in the file system implies non DB solution for geoprocessing
  - Storing in the DB implies geoprocessing in the DB

PostgreSQL/PostGIS

- **A secure/professional spatial database management system**
- **VERY robust and powerful geometric operations engine**
- **VERY large file capacity**
- **Multiple queries (users) at the same time**
- **Queries simple to write (SQL), no development needed**
  - shp2pgsql c:/temp/buffers.shp buffersTable | psql foundrydb
  - SELECT ST_Intersection(l.geom, b.geom)
    FROM lakesTable l, buffersTable b
    WHERE ST_Intersects(l.geom, b.geom)
- **Clean & secure storage**
- **Simple to deploy and scale**
- **Free!**
- **But missing raster capacities…**

PostGIS Raster

- **Open Source project adding support for rasters in PostGIS**
- **Loader based on GDAL –» support for MANY raster formats**
- **Takes nodata values into account, multiple bands, tiled and indexed (much faster), overviews**
- **SQL API**
  - get & set raster properties, georeference, convert to/from geometries, export to other raster formats
  - resample, reclass, reproject
  - compute statistics, intersect with geometries, map algebra
- **Almost transparent queries, familiar to PostGIS users**
  - raster2pgsql c:/temp/*.tif rastTable | psql foundrydb
  - SELECT ST_Intersection(a.rast, b.geom)
    FROM rastTable a, bufferTable b
    WHERE ST_Intersect(a.rast, b.geom)
- **2 years development involving many organizations**
- **Released with PostGIS 2.0 April 3th 2012**

Back on Our Requirements

1. **A good storage solution** ✓
2. **A robust and powerful raster/vector geoprocessing engine** ✓
3. **A web interface to**
   1. upload shapefiles in the store
   2. document administrators and users datasets
   3. build intersection queries
4. **Optionally a web mapping engine**

- **Capacity to develop a complete customizable geoweb application**
- **Ideally without (or with minimal) programming**

A Simple and Complete Metadata Profile

- **ISO 19115 is way too complex to maintain for small organizations**
  - +400 fields of information
  - VERY technical language
  - Designed for governmental organizations having a lot of resources
- **Feature documentation (all the attributes of a feature layer) is not part of ISO 19115 –» ISO 19110**
- **Our simpler subset (profile) of ISO 19115 and ISO 19110**
  - 29 mandatory fields, 32 optional
  - mix of ISO 19115 and ISO 19110 fields
  - vocabulary adapted for non GIS experts

GeoNetwork?

- **Web cataloging solution used as base for many SDI**
- **Store in many mtetadata standards (ISO 19115, ISO 19110, ISO 19119, ISO 19139, FGDC, Dublin Core)**
- **Serve in many protocols (CSW, Z39.50, GeoRSS, WebDAV)**
- **Search in and harvest from other catalogs**

GeoNetwork Opensource

Pros

- **Complete standard search/view/edit metadata solution**
- **Very good connection with other software**

Cons

- **Poor web usability, hard to customize**
- **Very hard to embed into an existing web application**
- **Poor data exploration mode, no facets**
- **Top down approach –» From experts for users**
- **Hardly relational when it should be**
  - no reuse of already entered entities other then was is provided by templates
- **Hard to implementation a profile not trivial**
- **No geoprocessing**

GeoNode?

- **Specifically developed to build a SDI**
- **Built with geo web development framework GeoDjango**

Pros

- **Very (Too?) simple metadata schema**
- **Bottom up approach –» From users for users :-)**

Cons

- **No support for ISO 19110 (only attribute names, no type, no description, no code list)**
- **Poor exploration mode, no facets**
- **No geoprocessing**
- **Very hard to embed into an existing web application**
- **Same problems with ESRI Geoportal Server**

A Geo CMS/Wiki?

- **All the services offered by a CMS or a Wiki**
  - Online site building
  - No HTML, no CSS, skin based
  - Easy form creation
  - Multi users and secure
  - Plenty of plugins for specialized functionalities
  - Many already integrate OpenLayers for mapping

- **What's missing?**
  - **GeoNetwork services**
    - for storing, publishing and editing metadata
  - **GeoServer (or MapServer) services**
    - for uploading data to the DB
    - for publishing them as geoweb services (WMS, WFS, WPS)
  - **Web Processing Services (WPS) interface**
    - configurable with CMS forms
    - plugable on many geoprocessing engines

- Geo services interfaces can be ***adapted*** to different ***classes*** of web application (like OpenLayers already does).
- ***GeoDjango*** is near from this but still requires ***Python development skills.***
- ***EasySDI*** is another attempt using ***Joomla.***

Our Direction…

- **Considering**
  - The complexity/time required to customize/simplify/extend GeoNetwork, GeoServer or GeoNode interface for our own application
  - The time required to develop a pluggable set of PHP drivers for GeoNetwork and/or GeoServer for a particular CSM or wiki
  - The fact that most services other than cataloging are not mandatory (mapping, WMP, WFS, WPS, CSW) for our application
- **We decided to develop the site using Semantic Mediawiki**
  - Semantic extension to the most used wiki in the world – Wikipedia – A database
  - Everything is customizable, no HTML, no CSS, no programming -> skins
  - Strong community building capacities, user management, security, change history
  - Many tools for search and exploration
  - VERY easy to build complex forms
  - Geo services will be implemented as they are needed
    - upload shapefiles to PostGIS
    - link a form to the geoprocessing engine
    - spatial search
    - export metadata to ISO 19139, 19110

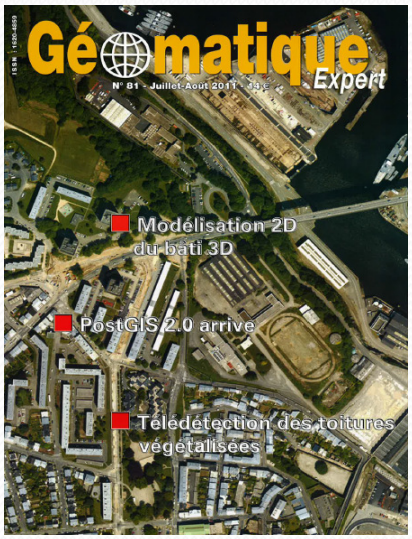
We should progressively end up with a full Web GIS in Mediawiki!

Conclusion

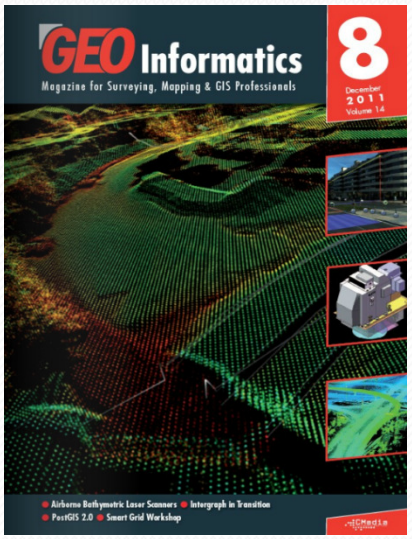
- **PostGIS with raster support provides a viable and robust storage solution offering geoprocessing engine**
- **There is not yet a easy/complete/flexible solution to implement a SDI**
  - GeoNetwork is for experts and is has numerous usability issues
  - GeoNode is hardly configurable and lack many functionalities
  - GeoDjango is a step toward the right direction but still requires developer skills
  - We need common geo services to be embeddable into existing web development tools like CMS or Wiki for easy specific application building
  - Not sophisticated/expert SDI solutions hard to adapt for our own specific needs
- **The trend is toward bottom up, social portals**
- **Still a lot of development to do to get easy tools**
- **In our case, all the features provided by a semantic wiki outmatch the services provided by GeoNetwork/GeoNode/ GeoDjango**
- **We will progressively tranform Semantic Mediawiki into a geo web site building tool and database**

Questions

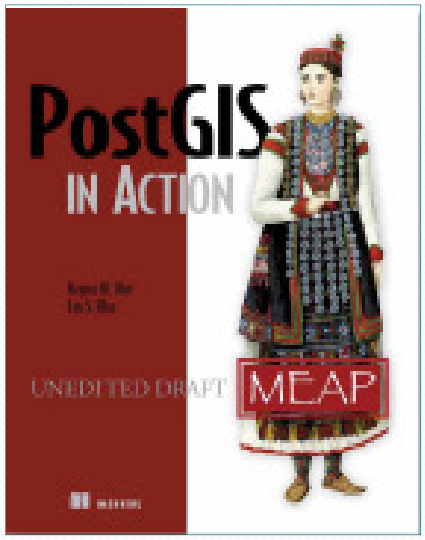
WKT Raster
Brazil FOSSGIS
June 2011



PostGIS 2.0 Arrive
Géomatique Expert
July 2011



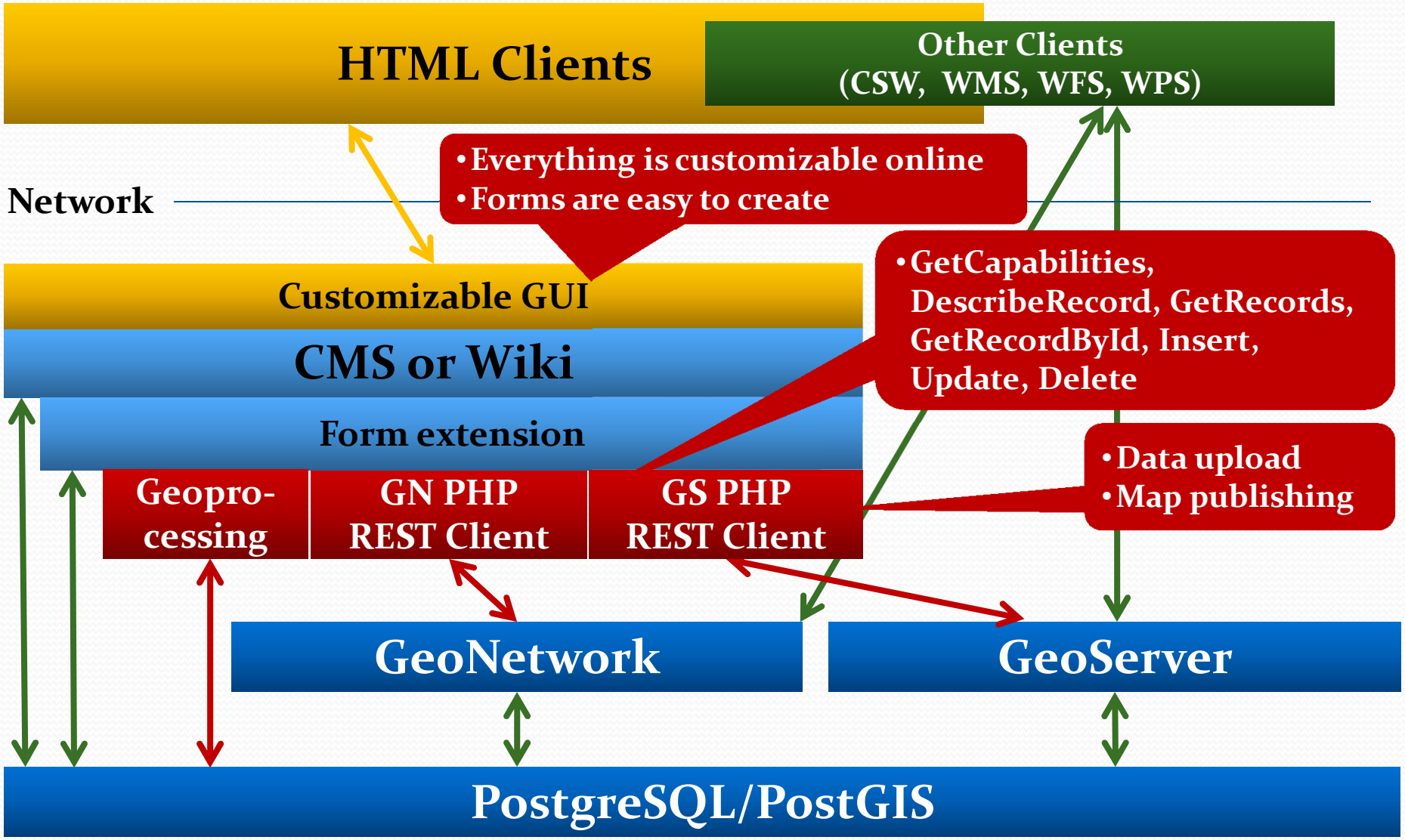
PostGIS In Action
GEO Informatics
December 2011



Chapter 13 on
PostGIS Raster
April 2011

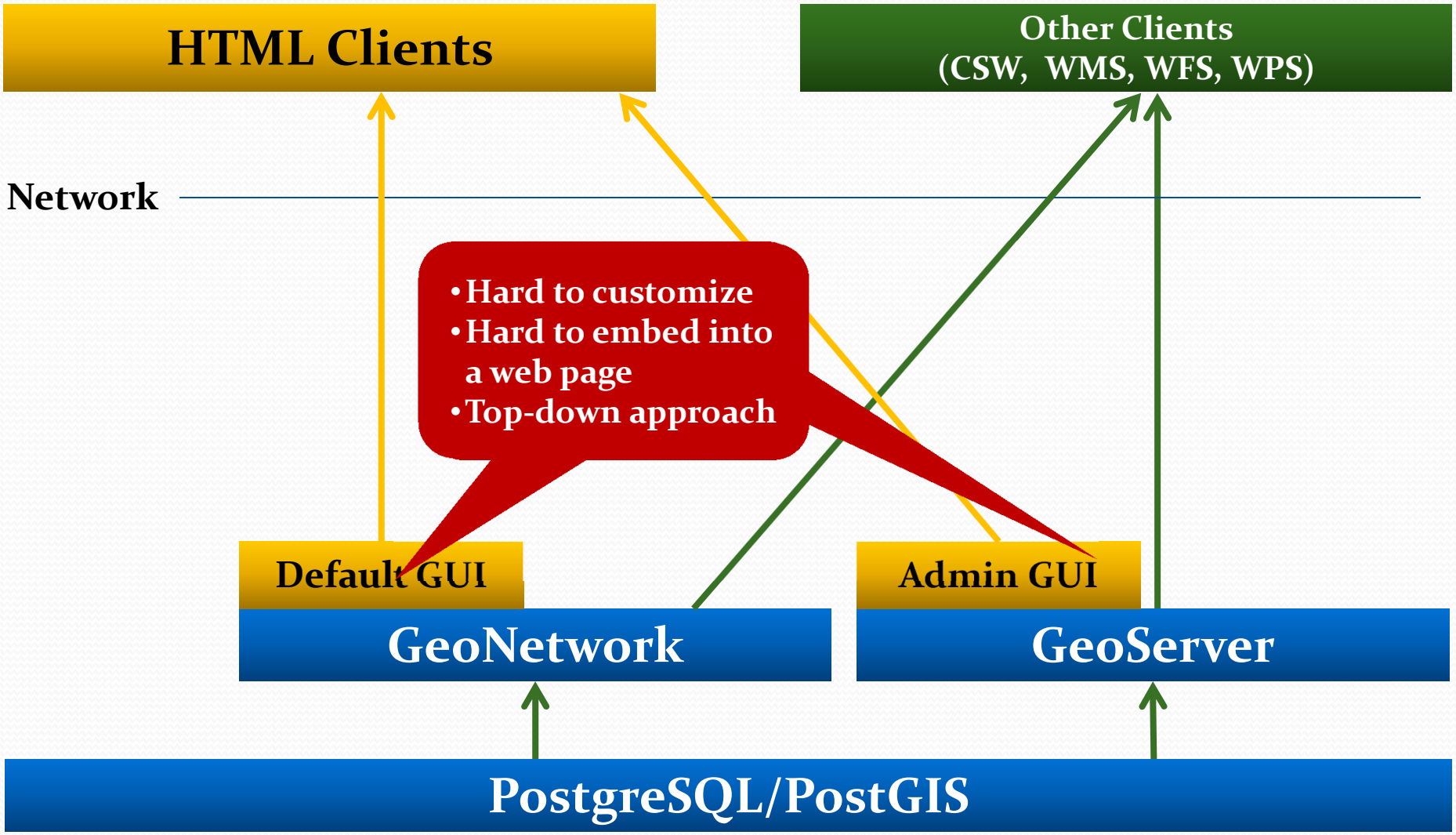
Thanks!

A CMS/Wiki Based SDI

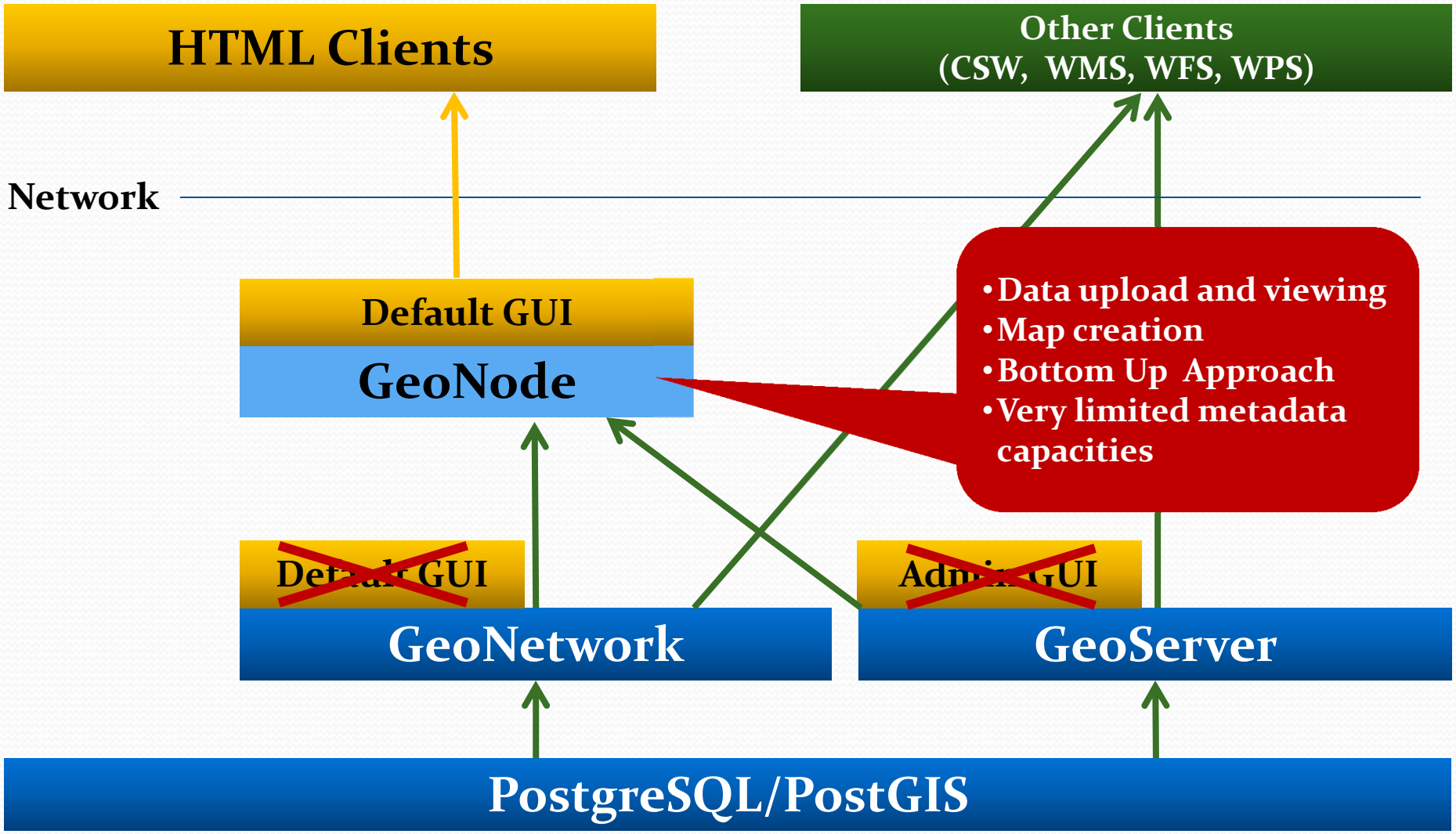


- **Some directions**
  - Implementing everything in Semantic MediaWiki
  - Integrating the functionalities of a SDI portal into an existing CMS or Wiki
  - Building everything using a geo web framework (GeoDjango)
  - Plug GeoNetwork as the documentation engine in GeoServer

A Simple GeoNetwork SDI



A GeoNode SDI



Do and do not when building an SDI

- **Focus too much on search and not enough on browse**
  - People often do not know what keywords to search for
  - Results in none or too many datasets
  - New users are in exploratory mode
  - Must be able to browse and restrict criteria (spatial, temporal, categorical, file type, etc…) at the same time (facets)
- **Focus too much on mapping**
  - Seeing the data is not what is the most important
  - Good facet (or criteria based) browsing is more important. Think about a library catalog: The same way a book reader is not a catalog, a map is not a catalog.
  - Too many buttons and feature, many do not work!
  - A preview is generally sufficient
- **Spliting products by mapsheet and returning each mapsheet when searching**
  - Once a product is found offer to cut it to a mapsheet or a specific area
- **Catalog base datasets, and derived datasets only to a certain point**
  - Criteria: how many applications can use this data
- **No users collaboration**
  - User knowledge is precious, let them document, comment, correct errors
- **Have strict policies on names and description**
  - Actually avoid names! Should be « Theme for location for time », Elevation for Canada
- **Propose preferred datasets and link to similar datasets**
- **Ideally everything should be centralised but licence restrictions and security make centralized data repository not wishable howeverlicensing does not prevent centralized metadata repository**
- **Be verbose and use a lot institutional jargon, not follow web usability rules**
  - Use all the screen!

Beyond the Traditionnal SDI

- **From a Download Oriented SDI**
- **To an Upload and Analysis SDI**

- **From an Organization Controlled SDI Paradigm**
  - Organization provide data, users consume data and results of analyses
- **To a User contributed SDI Paradigm**
  - Users consume, upload and analyse data
- **To a Social Network SDI Paradigm**
  - Users consume, upload and share data and analyses and about data and analyses