Workflow and Data Management in Spatial Studies: Just an extra document or something with intelligence?

Auteur: John Stuiver, Wageningen University
This can be quite a puzzle

Information and Communication needed
Information is:

- “Information is a source of learning. But unless it is organised, processed, and available to the right people in a format for decision making, it is a burden, not a benefit.” C William Pollard (1996)
Repeatability and Reproducibility

- **Repeatability** in this context is the *method* of how the work will be/was done.

- **Reproducibility** in this context is how the *work* is/was *implemented* and *carried out*.

New developments to support this:

→ **Workflow Management** to support this:

**Data Management Plan (DMP)**
Scientific Workflows

- “These are networks of analytical steps that may involve, e.g., data collection, database access, querying steps, data analysis, and mining steps, and many other steps including computationally intensive jobs on high performance cluster computers.”

Bertram Ludäscher et al. (Kepler project)
- [http://kepler.nasa.gov/](http://kepler.nasa.gov/)
Workflow Software Developments of Importance

- Enterprise Architect Sparx Systems
  - *Software ideas modeler*
- Workflow Management in ArcGIS 10 *ESRI*
- These developments are Based upon UML
  
  "The Unified Modeling Language (UML) is a graphical language for visualizing, specifying, constructing, and documenting the artifacts of a software-intensive system."
User demands

- Keep it simple!
  - **UML Activity model** is the most simple UML model. It is a **Dynamic Model** that describes how “activities” interact with given objects (*data*).

- Highly automated!
  - Graphical programming (data-action models) and documenting (Enterprise Architect market leader)
    - (Software Ideas Modeler – semi-open source)

- Must not confine user in finding a work solution!
  - Low level threshold in us (very difficult to achieve)

- Must have an added value!
  - For example automatically generate full metadata document
Why are (Scientific) workflow applications not available?

From literature two main reasons:

1. No framework exists to represent computational and none-computational workflows.

2. The ability of researchers to generate data in contemporary computing environments can quickly exceed their ability to track how it was created. (Trial and Error when using software approach)

Needed is a general framework
→ Does not exist (yet)
Framework proposal – Remember Geotools 😊
Project Application example
Overview processing and data involved

```
\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a001aConvertOriginalReceivedDataToProjectSourceData

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

Name

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

Calculate Value

Yrs

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

ps001LsYr

Copy Features

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a001ALsYr

Add Field (2)

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a001ALsYr (2)

Calculate Field (2)

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a001ALsYr199700 (2)

Copy Features (2)

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a001ALsYr%Yr.%s% (3)
```

PreProcessing Models that create Project Source and Component data for the project Paper 3 Jalal Samia

```
\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM.ps001andslides1

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM.ps001andslides1

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM.ps001andslides1

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM.ps001andslides1

\a0GDBData\GDBProject1Results.gdb\ps001andslides1

a002ComputeDerivativesDEM.ps001andslides1
```
Ordering -- Storing processes and data 1

From

- Add_weight_values_and_meaning_to_polygon_data
- Add_weights_etc_to_all_ps_data
- Aggregate_houses
- Compute_most_desired_light_rail_and_area
- Compute_preference_buffer_and_remove_houses_buildings
- Convert_poly_to_weight_grids
- Copy_project_infra_structure_to_temp_GDB_and_add_bufdis
- Create_Buffer_Distance_Table_and_give_buffer_values
- Create_source_grids
- Create_a_clipbox_feature_for_project
- Create_all_project_source_data_and_weight_tables
- Create_empty_target_weight_grid
- Create_proj_source_data_and_start_weight_tables_0
- Create_required_GDB_0
- Create_weight_grid
- CreateProject_clipbox_using_xy_minmax_in_RD_0
- Import_demography_to_GDBUsed_inIAD???

To

- IA_LightRailApplicationPublishVersion.tbx
- a_General_project_preprocessing_tools
  - IAa0_special_comp_scripts
    - a0stCreateClipboxExtentFeatureForProject
  - a_project_preprocessing_optimal
    - a01stCreateProjectSourceDataPlusFrequencyTablesTopoData
    - a02stMakeBasicReclassificationTable
    - a03stMakeDemographicData
    - a04stMakeProjectSuitabilityWeightTable
    - a05stv2MakeDEM
    - a06stSelectTrainStations
  - b_ComputeLROptimalVersion
    - b0st0ComputeBridgeSuitability
    - b1st0JoinSuitabilityWeightValuesToLanduseData
    - b2st0CreateSuitableZoneAroundMainSecondaryRoadsAndAggHouses
    - b3st0SlopeFillFlowAccumulationRestrictions
    - b4st0MakeSuitRasterUsingUpdate
    - b5st0ComputeCostPathCorridor5ResolutionAreaSources
  - d_Calculate_Percentage_within_Walking_Distance
    - d0stAlternativeWalkZoneCalculation
    - d1stv1ComputePercentageWithinWalkingDistanceUsingBuffer
    - d1stv2ComputePercentageWithinWalkingDistanceUsingAlternativeBuffer
    - d2stExampleForLoop
  - e_Compute_Cost_LightRail
    - e1stComputeLightRailCost
Ordering -- Storing processes and data

- From
- To
Aspects to consider at the start of a project

Data Management Plan (DMP) deals with:

- The **types of data, software and other materials produced** within a project;
- The **standards** to be used for **data and metadata format and content**;
- **Policies** for access, sharing, and ownership of data and products;
- **Plans for archiving** data, software and other project products, and for the preservation of access to them.
Aspects to consider at the start of a project

This DMP applies to all projects performed at the GRS chair group.

To whom this plan applies:

1. Ph.D candidates
2. GRS projects
3. Thesis and All projects at BSc and MSc level!
Aspects to consider at the start of a project

- **Basic elements of a DMP**
  1. Description of organizational context
  2. Short description of project
  3. Define data management roles
  4. Give an overview of project data (Source and Result), software choice(s), data size plus growth estimate
  5. Short term storage solutions
  6. Structure data and information
  7. Documentation and metadata
  8. Sharing and ownership
  9. Long term storage
Aspects to consider at the start of a project

- Data sources:
  - What metadata information is important in your case?

  Example FAO site:


- Metadata Style

  The style determines how metadata is viewed, exported, and validated, and which pages appear when editing metadata.

  ISO 19139 Metadata Implementation Specification
Aspects to consider at the start of a project

- Data and script sources:
- Do these data have meta-information?
  - If so, is this according to a meta-data standard organisation: ISO, CEN, NEN ...
  - ISO 19115 (formal description)
  - **ISO 19139 (technical implementation)**
  - Catalog service (CAT)
  - **Meta-data about services (ISO 19119)**
  - http://www.isotc211.org/
  - http://geostandards.geonovum.nl
  - DEMO
Aspects to consider at the start of a project

- Issues to check:
  - Geometric issues:
    - Single feature or multi-feature dataset?
    - Proper data format for analysis?
    - Shape files??? : Alternative??
  - Thematic issues
    - Attribute schema available?
      - Attribute names dataset correspond?
    - Percentage “Null Values” in attributes?
    - “Domain” attribute values known?
Aspects to consider in the project

- Data should be prepared for sharing
  - Team members should be able to access data at all times.
  - Data file names must describe content
    - Landuse_clip 😞
    - psa2LanduseProjectArea 😎
  - Attribute names must describe content
    - Comp_field1 😞
    - RatioAreaLengthParcel 😊
Aspects to consider in the project

- Aspects to be considered:
  - Give metadata to *all* project source and result data; both spatial data and tabular data
    - Use a standard!
      - Metadata editor available in ArcGIS software - ArcCatalog
  - Make sure attribute names explain content
  - **Use an export or exchange format that does not change anything of your end result!**
Aspects to consider in the project

- Use an export format that does not change anything of your end result!
  - Note Shape and DBF files can/will change attribute names!!
  - A good alternative is SQLlite
Aspects to consider at the end of a project

Must have a document describing your Geo-processing thread of end result data to guarantee reproducibility

- Bose, 2004
Further case experience and research necessary

Willing to give it a go?
Thank you for your time and patience
Typical Primary Questions in a Project

- What are the **project goals**?
- What **questions/actions** need to be **answered/carried out**?
- What are the **expected deliverables**?
- Which **methods/applications** are being used? How are these data made?
- How can I **monitor** the project **progress** and application **development**?
- What **data** is **needed** and **which data are relevant to save**?
- **To what extent** is the working method **repeatable**?
- **To what extent** are the project results **reproducible**?
- What **information/metadata** should be **available** during the project and at its completion?

Centre for Geo-Information
Workflow Software Developments of Importance

- The UML offers a standard way to write a system's blueprints, including conceptual, things such as processes and system functions as well as concrete things such as programming language statements, database schemas, and reusable software components."

- An extensive UML documentation is not requested during this course, HOWEVER......
Aspects to consider in the project

- **All** project results must have metadata; *both spatial and tabular data*

- Use a metadata standard
  - Metadata editor available in ArcGIS software – ArcCatalog
  - Also save presentation files to exactly reproduce a presentation
  - *Not in ArcGIS you can save a Map or Layer to a packages. When you unpack this packages also the presentation symbology is saved!*